大数据导论

关键技术与行业应用最佳实践

INTRODUCTION TO **BIG DATA**

深圳国泰安教育技术股份有限公司大数据事业部群
中科院深圳先进技术研究院—国泰安金融大数据研究中心　编著

揭开大数据的神秘面纱，全面解读大数据领域的

应用现状、原理、热门技术、前沿工具和解决方案。

清華大学出版社

大数据导论

关键技术与行业应用最佳实践

深圳国泰安教育技术股份有限公司大数据事业部群
中科院深圳先进技术研究院—国泰安金融大数据研究中心　编著

清華大学出版社
北　京

内容简介

本书全面阐释了大数据的概念、相关的技术和应用的现状，使读者对大数据的相关技术、应用和产业链能有一个比较清晰的认识。

全书共 11 章，主要内容包括大数据概论、数据组织存储技术、NoSQL、Hadoop 和 MapReduce、数据查询和分析高级技术、数据挖掘技术、数据分析语言 R、大数据用于预测和决策、大数据与市场营销、大数据应用案例、大数据应用主流解决方案等。

本书在内容的选择上进行了深入的思考，不论是大数据领域的初学者还是具备一定相关专业知识的读者都能从书中得到一定的收获或启发，同时，本书还适合高等院校的计算机相关专业的本专科生、研究生以及 IT 行业的从业人员，和所有对大数据感兴趣的人士阅读。

图书在版编目(CIP)数据

大数据导论：关键技术与行业应用最佳实践 / 深圳国泰安教育技术股份有限公司大数据事业部群，中科院深圳先进技术研究院——国泰安金融大数据研究中心　编著. —北京：清华大学出版社，2015

ISBN 978-7-302-39271-2

Ⅰ.①大… Ⅱ.①深… ②中… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2015）第 024762 号

责任编辑： 杨如林
装帧设计： 深圳国泰安教育技术股份有限公司
责任校对： 徐俊伟
责任印制： 杨　艳
出版发行： 清华大学出版社

网　　址：http://www.tup.com.cn，　http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　　邮　　编：100084
社 总 机：010-62770175　　　邮　　购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者：河北新华第一印刷有限责任公司
经　　销：全国新华书店
开　　本：190mm×260mm　　印　　张：24.25　　字　　数：590 千字
版　　次：2015 年 3 月第 1 版　　印　　次：2015 年 3 月第 1 次印刷
印　　数：1～4000
定　　价：57.00 元

产品编号：062431-01

编 委 会

总　序

大数据一词，最早出现于20世纪90年代。随着云计算和物联网的不断发展，大量数据源的出现导致了非结构化和半结构化数据的迅速增长，数据单位也由TB级别跨越到了ZB级别，大量信息源产生的这些数据已远远超越目前人力所能处理的范围，人们在思索如何对这些数据进行管理及使用时，逐渐探索出一个新的领域。

大数据的“大”不仅指其容量，还体现在多样性、处理速度和复杂度等方面。无论人们是否关注过，海量的数据已如决堤之洪流涌入人们的生活，大数据的时代已然到来了。可以目睹的是，大数据的激流已经给个人生活、企业经营乃至国家和社会的全面发展带来了新的机遇与挑战。2012年，世界经济论坛年会的重要议题之一是“大数据、大影响”；美国也开始从开放政府数据、开展关键技术研究和推动大数据应用三个方面来布局其大数据产业；2011年以来，中国计算机学会、中国通信学会先后成立了大数据委员会——研究大数据中的科学与工程问题；中国《“十二五”国家战略性新兴产业发展规划》也提出了支持海量数据存储、处理技术的研发与产业化……大数据正以不可抵挡之势席卷全球。

随着大数据技术和市场的快速发展，驾驭大数据的呼声渐涨，蕴含在大数据中的价值使得大数据已经成为IT信息产业中最具潜力的蓝海，这也使得学习及掌握国际前沿的大数据处理工具和解决方案中的核心技术显得十分迫切。从全球角度来看，对大数据的认识、研究和应用还都处于初期阶段，特别是对我国来说，大数据真正落地，还需要一个长期的过程。由于大数据领域的研究和分析方法综合了云计算、数据仓库、统计学、数据挖掘、机器学习、数据可视化等学科知识，因此编写一套系统的大数据技术与应用的丛书，作为学习和掌握大数据相关理论与方法的开端，无疑是一件意义非凡的事。

为了满足国内大数据领域学术界和产业界系统学习和掌握大数据理论及分析方法的迫切需要，同时也为解决国内大数据相关专业，尤其是本科生、研究生教材过于零散不够系统等问题，深圳国泰安教育技术股份有限公司大数据事业部群和中科院深圳先进技术研究院——国泰安金融大数据研究中心合作，组织该领域的专家、学者编写了“大数据技术与应用丛书”。这套丛书包括《大数据导论：关键技术与行业应用最佳实践》《Datawatch在各行业的应用》等。这套丛书是我们结合了大数据发展技术、发展现状，并调查了国内学术界和产业界的实际需求后精心编写的。希望这套丛书的出版，能为中国大数据学术界和产业界吸纳国内外先进的研究理念和研究方法，为国内大数据领域的学术研究、学术发展和实务运作提供

支持与帮助。

编写出版这套丛书是一项长期的工程，从2014年初开始策划编写这套丛书，包括咨询专家、选择书目、组织编写、校对内容、图书出版，从策划者、编写者到校对者和出版者都投入了大量心血和时间。在选择书目时，我们主要考虑所选书目要尽量保证理论体系的完整性，涵盖了大数据领域最新的理论研究和技术操作，内容编排循序渐进，方法详尽且操作步骤简单明了。

由于时间关系，书中难免存在不妥和疏漏之处，敬请读者给予批评指正。

深圳国泰安教育技术股份有限公司董事长、
上海交通大学教授、博导
陈工孟
2015年1月

前　言

随着云计算、物联网等技术的不断发展和应用，海量数据在生产经营、商务活动、社交生活等领域不断产生。世界处在信息时代，美国奥巴马政府将大数据提升到国家战略层面，启动了“大数据研究和发展计划”；企业将大数据作为提高自身竞争力的重要手段；最热的IT词汇中“大数据”必有一席之地……数据思维深刻影响着人们的工作和生活，这让人们真切地感受到大数据时代已经到来。

大数据即将带来一场颠覆性的革命，它将推动社会生产取得全面进步，助推政府、金融、医疗、教育、零售、制造业、能源和交通等行业产生根本性的变革。大数据是一个事关国家、社会发展全局的产业，围绕产业链的上下游，大数据将带动智能终端、服务器和信息服务业等产业发展，有效减少社会运行成本，提高社会和经济运行效率。

为了在信息时代立于不败之地，了解大数据相关的知识是必要的。本书全面阐释了大数据的概念、相关技术和应用现状，使读者对大数据的相关技术、应用和产业链有一个比较清晰的认识。本书适合高等院校计算机相关专业的本专科生、研究生、IT行业的从业人员和对大数据感兴趣的人士阅读。本书在内容的选择上进行了深入的研究，使得不论是大数据领域的初学者还是具备一定相关知识的专业人员都能从书中得到一定的收获和启发。

本书共11章，内容涵盖大数据的基本概念、关键技术和行业应用解决方案。对技术感兴趣的读者可以重点阅读第2～7章，这几章全面介绍了数据的存储和分析技术；对大数据应用感兴趣的读者可以优先阅读第8～10章；第11章对那些想了解大数据产业链的读者会很有帮助。

第1章简要介绍大数据的概念、大数据与商业智能的关系和大数据的相关技术及发展趋势；第2章介绍数据组织和存储的关键技术；第3章重点介绍了NoSQL；第4章介绍了Hadoop和MapReduce相关技术；第5章阐明了数据查询和分析技术，对常用的分析工具进行了介绍；第6章对数据挖掘技术的概念、挖掘算法和数据挖掘的发展趋势进行了分析；第7章重点介绍数据分析语言R；第8章论述大数据在预测和决策方面的作用，并阐述了商业和政府决策管理的机遇和挑战；第9章包括大数据与市场营销的联系和大数据时代的营销模式创新等问题；第10章简述了大数据在金融、医疗、互联网和影视等行业的应用案例；第11章主要介绍大数据产业链，介绍了新兴科技企业（如Cloudera、深圳国泰安教育技术股份有限公司等）和传统IT巨头（如IBM等）在大数据领域的主流解决方案。

新思想、新技术的不断涌现，推动着大数据的成熟与应用，也有效地推动着科技和社会

的进步。本书结合了深圳国泰安教育技术股份有限公司的产品和解决方案，为读者呈现了大数据领域的全景图。在成书过程中，我们参考了大量国内外学者的研究成果和业界的产品及解决方案，资料来源列在每章参考文献中，在此对各位学者和专业人士表示敬意和感谢！

由于作者水平和时间有限，书中难免存在疏漏和错误之处，恳请读者批评指正。

编　者

2015年1月

目　录

第1章　大数据概论

第2章　数据组织存储技术

第3章　NoSQL

第4章 Hadoop和MapReduce

第5章　数据查询和分析的高级技术

第6章　数据挖掘技术

第7章　数据分析语言R

第8章 大数据用于预测和决策

第9章　大数据与市场营销

第10章　大数据应用案例

第11章　大数据应用的主流解决方案

第1章
大数据概论

相信很多读者一定还记得2012年“双十一”光棍节淘宝的销售记录，当天销售额高达191亿人民币，相当于2012年全国前11个月社会消费品零售总额18.6万亿的5%。之后阿里巴巴董事局主席马云说，这是中国经济转型的一个标志，这一年可谓是电商的“井喷”之年。2013年 “双十一”光棍节，淘宝销售记录更是高达350.19亿元。自从2009年淘宝在11月11日开展“品牌商品5折优惠”活动以来，这一天的交易额由2009年的1亿元、2010年的9.36亿元、2011年的52亿元、2012年的191亿元，一直狂飙到2013年的350.19亿元，正式超越美国“网络星期一”，成为世界上最大的购物狂欢节。这一系列惊人的创举背后是什么？是大数据。它成就了阿里，也成就了中国的电商时代。

2011年5月，全球知名咨询公司麦肯锡（Mckinsey and Company）发布了《大数据：创新、竞争和生产力的下一个前沿领域》的报告，标志着大数据时代的到来。2012年世界经济论坛发布了《大数据：大影响》的报告，从金融服务、健康、教育、农业、医疗等多个领域阐述了大数据给世界经济社会发展带来的机会。

整个世界已经迎来了大数据时代。根据最新调查结果显示，2015年将会有近200亿个设备连接到互联网上，这些设备不仅有电脑，更有汽车、工厂设备、数字标牌等之前不可想象的设备。越来越多的智能终端设备给产业发展带来巨大的机遇。到2020年，人类产生的数据总量将达到40ZB，全球范围内服务器的数量将会增加10倍，由企业数据中心直接管理的数据量将会增加14倍，IT专业人员的数量将会增加1.5倍。许多权威人士认为这一数据大爆炸堪比新型石油，甚至是一种全新的资产类别。

1.1　什么是大数据

互联网、移动互联网、物联网、云计算的快速兴起，以及移动智能终端的快速发展，造成当前数据增长的速度比人类社会以往任何时候都要快。数据规模变得越来越大，内容越来越复杂，更新速度越来越快，数据特征的演化和发展催生出了一个新的概念——大数据。

最早引用的所谓“大数据”概念，可以追溯到Apache公司的开源项目Nutch。当时，把大数据描述为用来更新网络搜索索引以及需要同时进行批量处理和分析的大量数据集。其实早在1980年，著名的未来学家阿尔文·托夫勒便在《第三次浪潮》这本书中，极力赞扬大数据为“第三次浪潮的华彩乐章”。不过，大概从2009年开始，“大数据”才成为IT行业的流行

词汇。根据美国互联网数据中心的数据，互联网上的数据每年将呈现50%的增长，即每两年将会翻一番。而实际上，世界上90%以上的数据都是最近几年才产生的。除此之外，数据又并非单纯指人们在互联网上发布的信息，全世界的工业设备、交通工具、生活电器、移动终端上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化情况，这也产生了海量的数据信息。

1.1.1 大数据的概念

何谓大数据，目前业界还没有公认的说法。就其定义而言，大数据是一个较为抽象的概念，至今尚无确切、统一的定义，各方对“大数据”给出了10余种不同的定义，比较典型的有以下几种。

研究机构Gartner认为：大数据是指需要借助新的处理模式才能拥有更强的决策力、洞察发现力和流程优化能力的具有海量、多样化和高增长率等特点的信息资产。

麦肯锡的定义为：大数据是指在一定时间内无法用传统数据库软件工具采集、存储、管理和分析其内容的数据集合。

维基百科的定义是：大数据指的是需要处理的资料量规模巨大，无法在合理时间内，通过当前主流的软件工具撷取、管理、处理并整理的资料，它成为帮助企业经营决策的资讯。

IDC对大数据作出的定义为：大数据一般会涉及两种或两种以上的数据形式。它要收集超过100TB的数据，并且是高速、实时的数据流，或者是从小数据开始，但数据量每年会增长60%以上。

Gartner给出的是一个比较宏观的定义。首先对数据进行了描述，并在此基础上加入了处理此类型数据的一些特征，用这些特征来描述大数据；而维基百科中的定义缺乏精确性，常用软件工具的范畴难以界定；麦肯锡和IDC又只强调数据本身的量、种类和增长速度，属于狭义定义。从大数据的概念看，对大数据的概念界定各有各的看法。“大数据”这一提法具有明显的时代相对性，今天的大数据在未来可能就不一定是大数据，从业界普遍水平看是大数据，但对一些领先者来说或许已经习以为常了。

狭义的大数据，主要是指大数据的相关关键技术及其在各个领域中的应用，是指从各种各样类型的数据中，快速地获得有价值的信息的能力。一方面，狭义的大数据反映的是数据规模非常大，大到无法在一定时间内用一般性的常规软件工具对其内容进行抓取、管理和处理的数据集合；另一方面，狭义的大数据主要是指海量数据的获取、存储、管理、计算分析、挖掘与应用的全新技术体系。

广义上讲，大数据包括大数据技术、大数据工程、大数据科学和大数据应用等大数据相关的领域。即除了狭义的大数据之外，还包括大数据工程和大数据科学。大数据工程，是指大数据的规划建设运营管理的系统工程；大数据科学，主要关注大数据网络发展和运营过程中发现和验证大数据的规律及其与自然和社会活动之间的关系。对大数据进行广义分类是为了适应信息经济时代发展需要而产生的科学技术发展的趋势。

1.1.2 大数据的特征

IBM公司认为大数据具有3V特点，即规模性（Volume）、多样性（Variety）和实时性

（Velocity），但是这没有体现出大数据的巨大价值。以IDC为代表的业界则认为大数据具备4V特点，即在3V的基础上增加价值性（Value），表示大数据虽然价值总量高但其价值密度低。目前，大家公认的是大数据有四个基本特征：数据规模大、数据种类多、处理速度快及数据价值密度低，即所谓的4V特性，如图1.1所示。

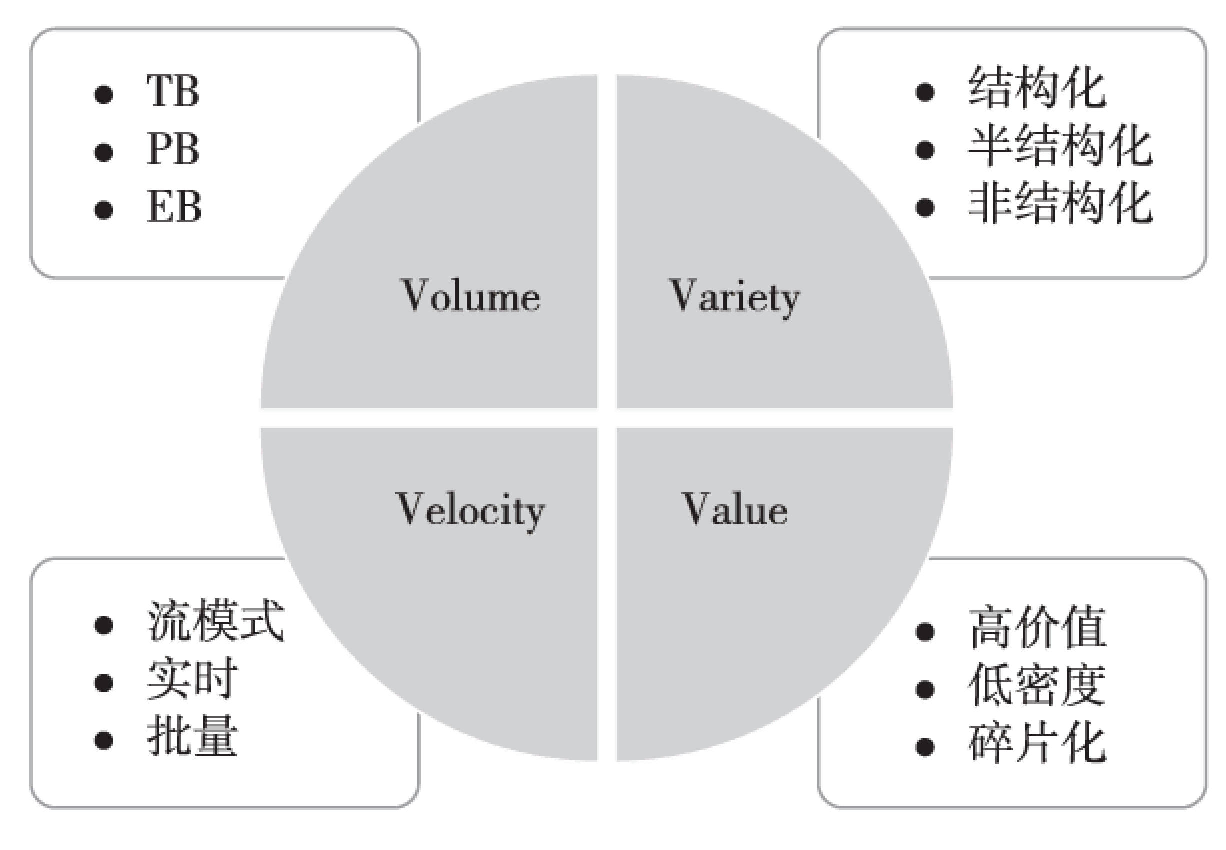


图1.1　大数据的4V特性

1. 数据规模大（Volume）

数据量大是大数据的基本属性，随着互联网技术的广泛应用，互联网的用户急剧增多，数据的获取、分享变得相当容易。在以前，也许只有少量的机构会付出大量的人力、财力成本，通过调查、取样的方法获取数据，而现在，普通用户也可以通过网络非常方便地获取数据。此外，用户的分享、点击、浏览都可以快速地产生大量数据，大数据已从TB级别跃升到PB级别。当然，随着技术的进步，这个数值还会不断变化。也许5年以后，只有EB级别的数据量才能够称得上是大数据了。

2. 数据种类多（Variety）

除了传统的销售、库存等数据外，现在企业所采集和分析的数据还包括像网站日志数据、呼叫中心通话记录、Twitter和Facebook等社交媒体中的文本数据，智能手机中内置的GPS（全球定位系统）所产生的位置信息、时刻生成的传感器数据等。数据类型不仅包括传统的关系数据类型，也包括未加工的、半结构化和非结构化的信息，例如以网页、文档、E-mail、视频、音频等形式存在的数据。

3. 处理速度快（Velocity）

数据产生和更新的频率也是衡量大数据的一个重要特征。1秒定律，这是大数据与传统数据挖掘相区别的最显著特征。例如全国用户每天产生和更新的微博、微信和股票信息等数据，随时都在传输，这就要求处理数据的速度必须要快。

4. 数据价值密度低（Value）

数据量在呈现几何级数增长的同时，这些海量数据背后隐藏的有用信息却没有呈现出相应比例的增长，反而是获取有用信息的难度不断加大。例如，现在很多地方安装的监控使得

相关部门可以获得连续的监控视频信息，这些视频信息产生了大量数据，但是，有用的数据可能仅有一、两秒钟。因此，大数据的4V特征不仅仅表达了数据量大，而且在对大数据的分析上也将更加复杂，更看中速度与时效。

1.1.3 大数据的产生

大数据的产生是计算机和网络通信技术（ICT）被广泛运用的必然结果，特别是互联网、移动互联网、物联网、云计算、社交网络等新一代信息技术的发展，起到了促进的作用，它使数据的产生方式发生了四大变化：首先，数据的产生由企业内部向企业外部扩展；第二，数据的产生由Web 1.0向Web 2.0扩展；第三，数据的产生由互联网向移动互联网扩展；最后，数据的产生由计算机或互联网（IT）向物联网（IOT）扩展。这四个方面的变化，让数据产生的源头呈几何级数地增长，数据量更是呈现大幅度地快速增加。

1. 数据的产生由企业内部向企业外部扩展

由企业内部的办公自动化（OA）、企业资源计划（ERP）、物料需求计划（MRP）等业务以及管理和决策分析系统所产生的数据，主要被存储在关系型数据库中。内部数据是企业内最成熟并且被熟知的数据，这些数据已经通过多年的主数据管理（MDM）、ERP、OA、MRP、数据仓库（DW）、商业智能（BI）和其他相关应用的积累，实现了内部数据的收集、清洗、集成、结构化和标准化处理，可以为企业管理决策提供支持与帮助。对于商业企业而言，信息化的运用环境在发生着变化，其外部数据也迅速扩展。企业应用、互联网应用和移动互联网应用之间的融合越来越快，企业需要通过互联网来联系外部供应商、服务客户，联系上下游的合作伙伴，并在互联网上实现电子商务和电子采购的交易和结算。企业需要开通微博、微信、QQ、博客等社交网络来进行网络营销、品牌建设和客户关怀。把电子标签贴在企业的产品上，在制造、供应链和物流的全过程中进行及时跟踪和反馈，必将有更多来自企业外部的数据被产生出来。表1.1所示为企业内外部数据产生的源头、规模及存储情况。

表1.1 企业内外部数据的产生

	企业内部数据	企业外部数据
企业应用	ERP、CRM、MES、SCADA、OA、专业业务系统、传感器	电子商务、电子采购、知识管理、呼叫中心、企业微博、企业微信、RFID、传感器
数据规模	TB级	PB级
数据存储	关系型数据库、数据仓库	各种格式的文档

2. 数据的产生从Web 1. 0向Web 2. 0，从互联网向移动互联网扩展

随着社交网络的迅速发展，互联网进入了Web 2.0时代，个人从数据的使用者，变成了数据的制造者，数据规模不断地扩张，每时每刻都在产生着大量的新数据。例如，从全球统计数据的角度来看，全球每分钟发送290万封电子邮件，电子商务公司亚马逊每秒钟将产生72.9笔商品订单，每分钟会有20个小时的视频上传到视频分享网站YouTube，谷歌每天需要处理24PB的数据，Twitter上每天发布5千万条信息，每个家庭每天消费的数据有375MB，每个月网民在Facebook上要花费7千亿分钟……

从中国的统计数据来看，数据规模也十分巨大。淘宝网会员超过了5亿，在线商品数超

过了8.8亿，每天产生的交易有数千万笔，产生约20TB的数据；目前百度拥有的数据总量接近1000PB，存储网页的数量接近1万亿页，每天大约要处理60亿次的搜索请求，产生几十PB的数据；新浪微博每天有数十亿外部网页和API接口访问需求，服务器群在晚上高峰期每秒要接受100万个以上的响应请求。

3. 数据的产生由互联网向移动互联网扩展

移动互联网的发展让更多的使用者成为数据的制造者。据统计，全球每个月移动互联网的使用者发送和接收的数据高达1.3EB。在中国，仅中国联通用户上网记录条数为83万条/秒，即一万亿条/月，对应数据量为300TB/月，或3.6PB/年。

4. 数据的产生从计算机/互联网（IT）向物联网（IOT）扩展

随着传感器、视频、RFID和智能设备等技术的发展，音频、视频、机器对讲机器、RFID、人机交互、物联网和传感器等数据大量产生，其数据量更是巨大。根据国际知名市场研究公司IDC公布的数据，在2005年仅机器对机器产生的数据就占全世界数据总量的11%，预计到2020年这一数值将可能增加到数据总量的42%。思科（Cisco）公司预测，仅移动设备的数据总流量在2015年就将达到每月6.3EB的规模。

1.1.4 数据的量级

数据规模的大小是用计算机存储容量的单位来计算的，最基本的单位是字节（Byte）。每一级按照千分位递增，最小的基本单位是Byte，按顺序所有单位依次为：Byte、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB。它们按照进率1024（2的十次方）来计算。

1KB= 1 024 Bytes

1MB= 1 024 KB = 1 048 576 Bytes

1GB= 1 024 MB = 1 048 576 KB

1TB= 1 024 GB = 1 048 576 MB

1PB= 1 024 TB = 1 048 576 GB

1EB= 1 024 PB = 1 048 576 TB

1ZB= 1 024 EB = 1 048 576 PB

1YB= 1 024 ZB = 1 048 576 EB

1BB= 1 024 YB = 1 048 576 ZB

1NB= 1 024 BB = 1 048 576 YB

1DB= 1 024 NB = 1 048 576 BB

巨著《红楼梦》含标点87万字（不含标点853 509字），每个汉字占两个字节，则1汉字=16bit = 2 × 8位=2bytes，以计算机单位换算，1GB 约等于671部《红楼梦》，1TB约等于631 903部，1PB 约等于647 068 911部。再以互联网为例，一天当中，在互联网上产生的全部内容可以刻满1.68亿张DVD；发出的邮件有2 940亿封之多；发出的社区帖子多达200万个，相当于《时代》杂志770年的文字量……截止到2012年，数据量已经从TB(1 024GB=1TB)级别跃升到PB(1 024TB = 1PB)、EB(1 024PB=1EB)乃至ZB(1 024EB=1ZB)级别。国际著名市场研究公

司IDC的研究结果表明，2008年全球产生的数据量高达1.82ZB，相当于全球每人产生200GB以上的数据。到2012年为止，人类生产制造的所有印刷材料的数据量是200PB，全人类历史上说过的所有话的数据量大约是5EB。IBM的研究称，整个人类文明所获得的全部数据之中，有90%是过去两年内产生的。到2020年，全世界所产生的数据规模将达到今天的44倍。

1.1.5 大数据的数据类型

大数据不仅仅体现在数量大，也体现在数据类型多。

1. 按照数据结构分类

按照数据结构分，数据可分为结构化数据与非结构化数据。非结构化数据又包含半结构化数据和无结构的数据。结构化数据通常存储在数据库中，可以用二维表结构来逻辑表达实现的数据。相对于结构化数据而言，非结构化数据是指不能用二维表结构来表现的数据，包括各种格式的办公文档、图片、图像、文本、HTML文档、XML文档，各类报表、音频和视频信息等。

（1）结构化数据

结构化数据的特点是在任何一列数据不可以再细分，并且任何一列数据都具有相同的数据类型。所有关系型数据库（如SQL Server、Oracle、MySQL、DB2等）中的数据全部为结构化数据。关系型数据库存储的结构化数据示例如表1.2所示。

表1.2 结构化数据示例

学 号	学生姓名	科 目	成 绩
20090903	张伟	数学	90
20090702	李东	英语	88

（2）半结构化数据

半结构化数据是处于完全结构化数据和完全无结构的数据之间的数据，这种数据类型的格式一般较为规范，都是纯文本数据，可以通过某种特定的方式解析得到每项数据。最常见的半结构化数据是日志数据、采用XML与JSON等格式的数据，每条记录可能都会有预先定义的规范，但是每条记录包含的信息可能不尽相同；也可能会有不同的字段数，包含不同的字段名、字段类型或者包含着嵌套的格式等。这类数据一般都是以纯文本的格式输出，管理维护相对而言较为方便。但是，在需要使用这些数据（如采集、查询、分析数据）时，可能需要先对这些数据格式进行相应地转换或解码。

下面是一个XML文档的示例。

```
<?xml version="2.0"?>
<Order>
<product xmlns="http://market">
<Title>The Joshua Tree</Title>
```

（3）无结构的非结构化数据

无结构的数据是指那些非纯文本类型的数据，这类数据没有固定的标准格式，无法直接解析出其相应的值。常见的无结构化数据有网页、文本文档、多媒体（声音、图像与视频

等）。这类数据不容易收集和管理，甚至是无法直接查询和分析，所以对这类数据需要使用一些不同的处理方式。

2. 按照产生主体方式分类

（1）最里层。由少数企业应用而产生的数据。

- 关系型数据库中的数据。
- 数据仓库中的数据。

（2）次外层。大量个人产生的数据。

- 社交媒体，如微博、QQ、微信、Facebook、Twitter等产生的大量文字、图片和视频数据。
- 企业应用的相关评论数据。
- 电子商务在线交易、供应商交易的日志数据。

（3）最外层。由巨量机器产生的数据。

- 应用服务器日志（Web站点、游戏）。
- 传感器数据（天气、水、智能电网）。
- 图像和视频。
- RFID、二维码或者条形码扫描的数据。

图1.2所示为不同的大数据主题示意图。

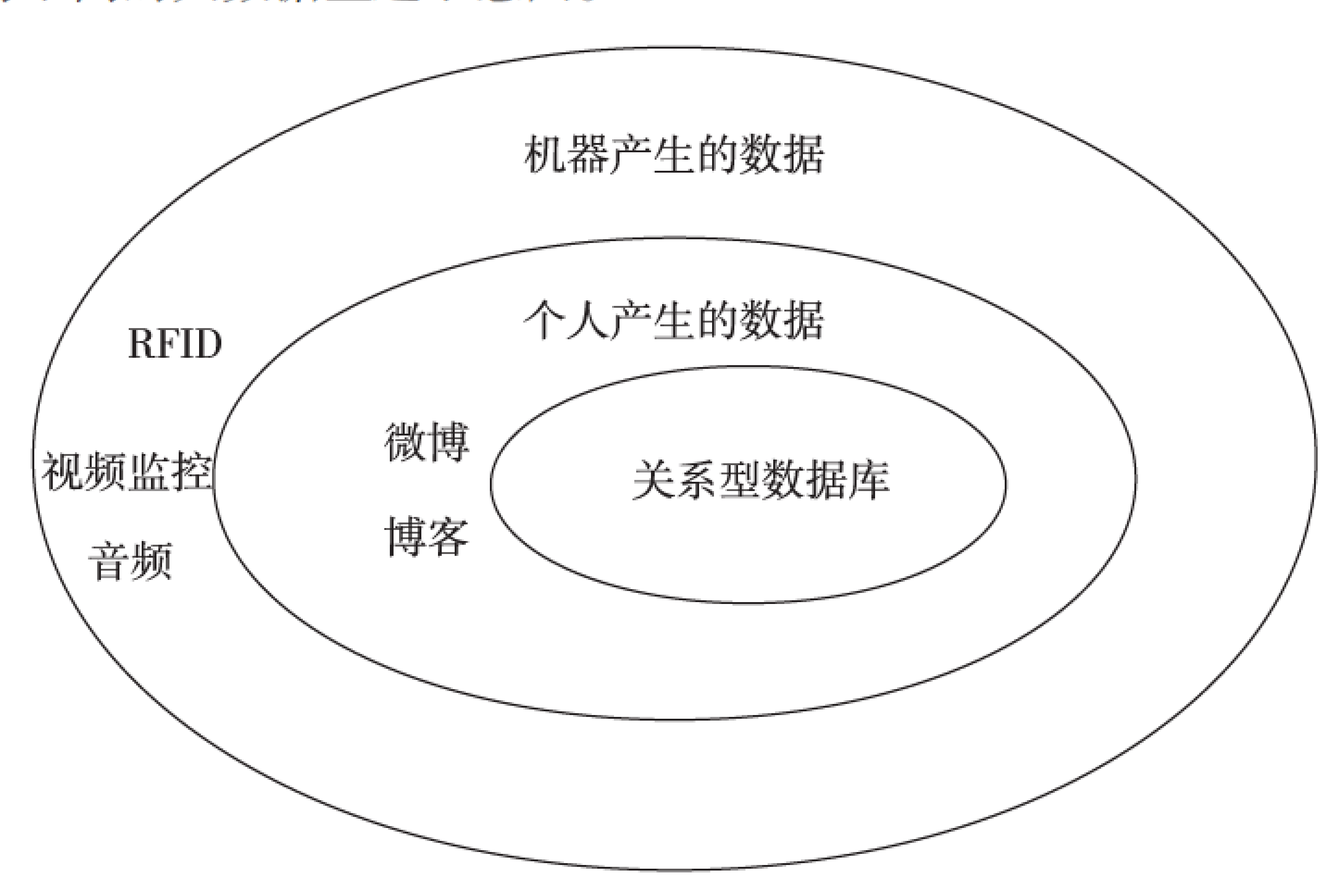


图1.2　不同的大数据主题

3. 按照数据产生作用的方式分类

按照数据产生作用的方式分类，可分为交易数据和交互数据。

交易数据是指来自电子商务或者企业应用中的数据，包括ERP、企业对企业（B2B）、企业对个人（B2C）、个人对个人（C2C）、线上线下（O2O）、团购等系统。这些数据存储在关系型数据库和数据仓库中，可以执行联机分析处理（OLAP）和联机事务处理（OLTP）。这些数据的复杂性和规模一直都在不断地增加。

交互数据指来自相互作用的社交网络中的数据，包括机器交互（设备生成交互）和社交

媒体交互（人为生成交互）的新型数据。

这两类数据的有效融合将是大势所趋。大数据应用要有效集成这两类数据，并在此基础上，实现对这些数据的处理和分析。

1.1.6 大数据的潜在价值

大数据的潜在价值可以通过数据结构的复杂性和关联性体现出来。当提到大数据时，我们最先想到的一定是其体量大，但是体量大的数据如果仅是简单的数据堆砌，或者仅是对单一类型数据的记录，那么这种重复性高、结构简单的数据还不能称之为大数据。例如，在一个购物商场内，商品种类有上千种，每种商品又有来自不同公司的产品，再加上购物、休闲、娱乐、餐饮等信息，则它拥有的数据就能从各个维度反映出顾客的行为特征，从而蕴含更大的数据价值。

大数据潜在价值的另一个体现是其关联性。大数据的重要来源之一是互联网行业。随着移动互联网的发展及互联网普及率的提升，网民上网行为呈现出跨网站、跨终端、跨平台等特点，用户数据不仅包括人与人交流产生的数据，还包括人机交互及机器与机器间通信产生的数据。这些数据之间如果没有较明显的逻辑关系和确定的关联关系，则数据价值的挖掘就会变得相当困难，同时数据价值也相应要低很多。所以数据之间的逻辑性和关联性也是数据潜在价值的蕴藏点。

大数据潜在价值的实现包括三个层次，社会领域、行业领域及企业发展领域。大数据最终需要解决的问题主要集中在这样三个层面上：一是宏观层面，主要是应用于社会领域，如智慧交通、智慧城市和灾难预警等；二是中观层面，主要表现在提升行业生产率水平、促进行业的融合发展以及促进行业内商业模式的变革等；三是微观层面，主要表现在促进客户服务水平的提升、企业流程的创新、内部运营成本的降低及供应链的协调和改善等。

1.1.7 大数据的挑战

大数据在带来巨大的潜在价值的同时，在业务视角、技术架构、管理策略等方面，由于存在差异性而形成了挑战。

1. 业务视角不同带来的挑战

在大数据未出现之前，企业通过对内部ERP、客户关系管理（CRM）、供应链管理（SCM）、商业智能（BI）等信息系统的建设，建立了高效的企业内部统计报表、仪表盘等决策分析工具，这些管理系统在企业业务敏捷决策方面发挥了很大作用。但是，这些数据分析只反映了冰山一角，因为报表和仪表盘其实是“残缺”的，是不全面的，更多潜在的有价值的信息往往被企业所忽略。在大数据时代，企业业务部门必须改变他们看数据的视角，要更加重视和利用以往被忽视的数据，如交易日志、客户反馈与社交网络等。这种转变需要一个接受的过程，但已经实现这种转变的企业则从中收获颇丰。据有关统计数据，电子商务企业亚马逊三分之一的收入来源于基于大数据相似度分析的推荐系统的贡献；花旗银行新产品的创意很大程度上来自于从各个渠道收集到的客户反馈数据。因此，在大数据时代，业务部门需要以更新的视角来面对大数据，接受和利用好大数据，以创造出更大的业务价值。

2. 技术架构不同带来的挑战

传统的结构化查询语言（SQL）和关系型数据库（RDBMS）在面对大数据时，已经显得力不从心，刺激性价比更高的数据计算、存储技术和工具不断地涌现。对于已经熟练掌握和使用传统技术的企业信息技术人员来说，接受、学习和掌握这些新的技术与工具需要一个过程，内心认为现在的技术和工具已足够好，对新技术会产生一种排斥心理，怀疑它只是一个新的噱头，同时新技术本身的不成熟性、复杂性和用户不友好性也会加深这种印象。应该看到的是大数据时代的技术变革已经不可逆转，企业必须积极迎接这种挑战，以包容的方式迎接新技术，以集成的方式实现新老系统的整合。

3. 管理策略不同带来的挑战

大容量和多种类的大数据处理将带来企业信息基础设施的重大变革，也在企业信息技术管理、服务、投资和信息安全治理等方面带来了新的挑战。像如何利用私有云、公有云等服务来实现企业内、外部数据的处理和分析，对大数据架构应该采取什么样的管理和投资模式，对大数据可能涉及的数据隐私应当如何保护……这些都是企业应用大数据需要面对的挑战。

1.2　大数据与商业智能

1.2.1　商业智能的概念

商业智能（Business Intelligence，简称BI），又称商务智能或商业智慧，其概念于1996年由Gartner Group提出。Gartner Group将商业智能定义为：商业智能是描述一系列的概念和方法，通过应用基于事实的支持决策系统来辅助商业决策的制定和实施。商业智能技术提供使企业迅速计算分析数据的技术和方法，包括收集、组织、管理和分析数据，并将这些数据转化为有用的信息，然后分发到企业各处。不过，目前公认的商业智能的定义是指：企业在收集、组织、管理和分析结构化与非结构化的数据和信息时，使用现代信息技术，使商务决策水平得以提升，商务知识和见解得以创造和增加，并且能够帮助企业完善商务流程，采取更有效的商务行动，提升各方面商务绩效，提高综合竞争力的智慧和能力。

商业智能是一系列技术、方法和软件的总称，其最终目的是提高企业运营性能以及增加企业商业利润。对于商业智能这个概念的正确理解，应从四个层面展开[①]。商业智能的转化如图1.3所示。

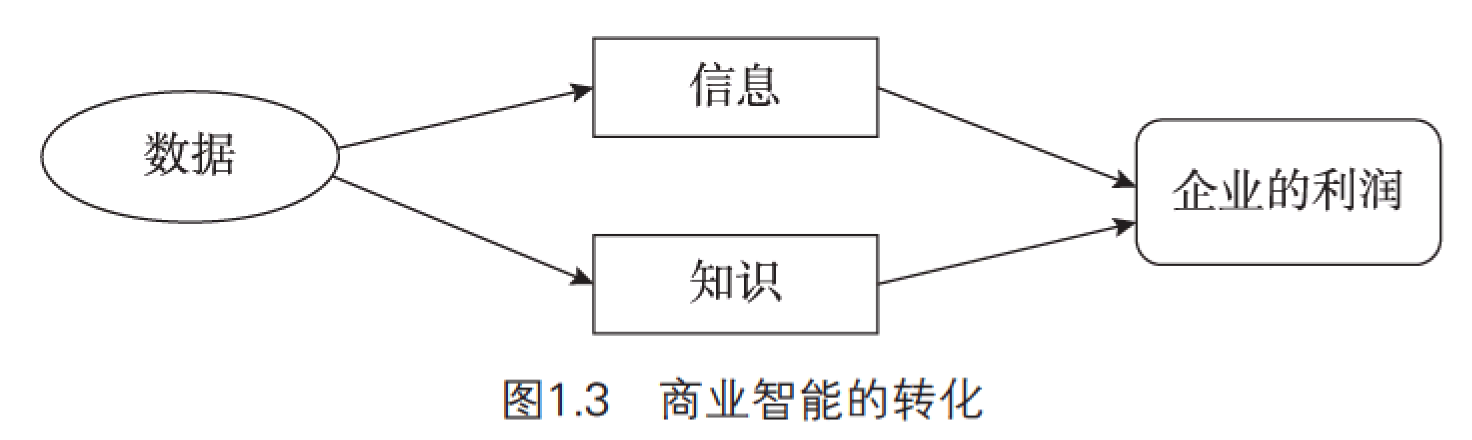


图1.3　商业智能的转化

① http://wenku.baidu.com/view/78088e6aaf1ffc4ffe47ac37.html.

第一，信息系统层面。它是商业智能系统（BI System）的物理基础，是一个面向特定应用领域的信息系统平台，一个独立的软件工具，具有非常强大的决策分析能力。

第二，数据分析层面。商业智能是一系列具有计算、分析功能的工具、算法或模型的总称。在数据分析层面，首先是获取数据，获取与所关心主题有关的高质量的数据或信息，然后自动或人工参与使用具有分析功能的算法、工具或模型，其间包括分析信息、得出结论、形成假设与验证假设等一系列的过程。

第三，知识发现层面。它与数据分析层面一样，也是一系列工具、算法或模型的总称。这一层面可以直接将信息转变成知识，或者是把数据转变成信息后，借助于大数据分析挖掘技术发现信息背后隐藏的东西，然后将信息转变成知识。

第四，战略层面。这一层面主要是将知识或信息应用在改善运营能力和提高决策能力以及企业建模等上面。商业智能的战略层面是提高企业决策能力，是通过利用应用假设或经验以及一个或多个数据源的信息所形成的一组方法、概念和过程的集合。它通过获取、组织、管理和分析数据，将数据和信息提供给贯穿企业组织的各类人员，使得企业的决策能力得以提高。

1.2.2 商业智能的架构体系

商业智能所涉及到的数据包括来自企业业务系统的订单、交易账目、库存、客户和供应商资料及来自企业外部即企业所处行业和竞争对手的数据，以及来自企业所处环境的其他外部的各种数据。商业智能所辅助的业务经营决策既可以是操作层面的，也可以是战术层和战略层的决策。要将数据转化为知识，需要利用数据仓库（DW）、联机分析处理（OLAP）工具和数据挖掘（DM）等技术。因此，从技术层面上来讲，商业智能并不是基础技术或者是产品技术，而是数据仓库、联机分析处理和数据挖掘等相关技术走向商业应用后形成的一种应用技术，其系统架构如图1.4所示。

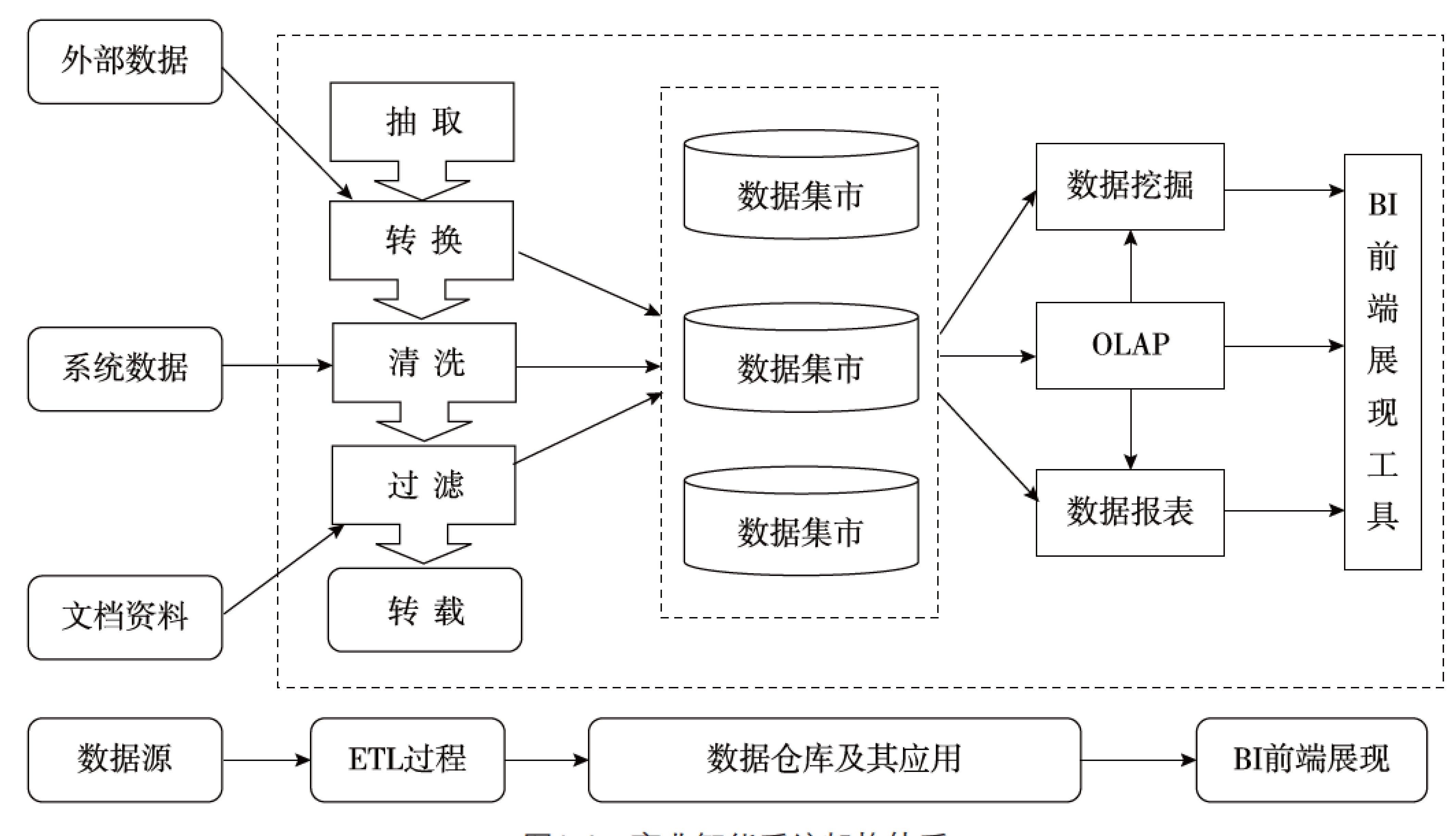


图1.4　商业智能系统架构体系

从图1.4中可以看到，实现商业智能应用有四个非常关键的环节，包括数据源、ETL过程、数据仓库及其应用和BI前端展现。

（1）数据源

数据仓库系统的数据来源主要是外部的操作性应用系统以及内部的业务系统。这些数据源包括数据的业务含义和业务规则，表达业务数据的表、字段、视图、列和索引。

（2）ETL过程

即抽取（Extraction）、转换（Transformation）和装载（Load）过程。ETL过程负责将业务系统中各种外部数据、关系型数据、遗留数据和其他相关数据经过清洗、转化和整理后放进中心数据仓库。

（3）数据仓库及其应用

数据仓库是商业智能系统的基础，是面向主题的、稳定的、集成的和随时间不断变化的数据集合。通过联机在线分析处理，可以对数据仓库中的多维数据实行钻取、切片以及旋转等分析操作，及时地完成决策支持所需要的查询及报表。通过数据挖掘，可以挖掘出数据背后隐藏的知识或信息。通过关联分析、聚类分析和判别分析等方法建立分析模型，来预测企业未来发展趋势以及将要面临的问题。

（4）BI前端展现

在海量数据和分析手段不断增多的情况下，BI前端展现的主要功能是保障系统分析结果的可视化。一般认为，数据仓库、联机在线分析和数据挖掘技术是商业智能的三大核心技术。决策者通过正确运用商业智能技术，将使用结果加以反馈。通过反馈，可以暴露出潜在的问题，同时，也可以根据情况变化，表达新的需求，提高商业智能流程内在质量。

商业智能为特定的应用系统（如客户关系管理CRM、企业资源计划ERP与供应链管理SCM）的数据环境和决策分析提供支持，在企业制订战略和决策时提供良好的支持。当面对特定应用的特定战略和决策问题时，商业智能能够从数据准备做起，建立或虚拟一个集成的数据环境，以集成的数据环境为基础，利用科学的决策分析工具，并通过数据分析、知识发现等过程为战略制订、决策分析、最终解释、执行分析和发现结果整个过程提供支持。在这个过程中，集成的数据环境和决策分析工具起了非常重要的作用。

1.2.3　商业智能的核心技术

商业智能实质上是数据转化为信息的过程，这一过程也可称为信息供应链，其目的是把初始的操作型数据变成决策所使用的商务信息。在这一过程中，数据集成工具执行原数据的清洗、格式转化和合并计算等功能；数据存储过程建立数据存储模型，存储企业统一的数据视图，为商业智能系统的应用提供基础数据；数据分析工具一般包括联机分析处理、统计分析工具、数据挖掘工具以及其他人工智能工具，这些工具结合商业处理规则，为决策者提供决策辅助信息。从商业智能系统建立的技术角度来看，构建一个完整的商业智能系统涉及到以下三种核心技术。

1. 数据仓库技术

在20世纪80年代中期出现了数据仓库技术。此技术被数据仓库创始人之一W・H・Inmon

定义为“数据仓库是一个面向主题的、集成的、稳定的和包含历史数据的数据集合，它用于支持管理中的决策制定过程”。数据仓库系统是对数据的处理技术的集成，而商业智能系统的核心是解决商业问题，它把数据处理技术与商务规则相结合以提高商业利润，减少市场运营风险，是对数据仓库技术、决策处理技术和商业运营规则的集合。数据仓库与传统数据库存储的最大区别在于，数据库用于处理企业日常事务，而数据仓库则主要用于处理商务运营决策。数据仓库建立的目的在于在不影响日常操作处理的前提下对业务信息进行分析以辅助企业决策，为决策支持系统提供应用基础。因此数据库与数据仓库是应用于企业运营过程中不同目的的两种数据管理系统。数据的存储技术是数据仓库技术的核心内容，在数据仓库中被集成的数据常常以星型模式展现出来，即以事实表—维表结构来组织数据。事实表也称为主表，包括商务活动定量的或实际的数据，这种数据是可以用数字来度量的，由多行和多列组成；维表又称为辅助表，一般比较小，是反映商业活动中某个维的描述性的数据。事实表和维表通过关系进行连接，这样的组织方式被称为多维数据存储的星型模式。在扩展的星型模式中，维表本身还能够包括维表，这样在组成的数据仓库中包含了商务事实的物理存储模式。

2. 数据挖掘技术

数据挖掘主要用于从大量的数据中发现或挖掘隐藏于其背后的规律或数据之间的关系，它通常采用机器自动识别的方式，不需要太多的人工干预。采用数据挖掘技术，可以为用户的决策分析提供自动化的、智能的辅助手段，该项技术在金融保险业、零售业、医疗行业等多个领域都可以得到很好的应用。在数据挖掘技术中常用的数据模型主要有：

① 分类模型，根据商业数据的属性将数据分配到不同的组别中。

② 关联模型，主要描述一组数据项目的密切度和关系。

③ 顺序模型，主要用于分析数据仓库中的某类同时间与之相关的数据，并发现某一时间段内数据的相关处理模型。顺序模型可以看成是一种特定的关联模型，它在关联模型中增加了时间属性。

④ 聚簇模型，当要分析的数据缺乏描述性信息，或者是无法组织成任何分类模式时，可以采用聚簇模型。聚簇模型按照某种相近程度度量方法，将用户数据分成互不相同的一些组。组中的数据相近，组之间的数据相差较大。聚簇模型的核心是将某些明显的相近程度测量方法转换成定量测量方法。

3. 联机分析处理

联机分析处理（On-Line Analysis Processing，简写为OLAP）的概念最早是由关系数据库之父爱德华·库德（E·F·Codd）博士于1993年提出的，是一种用于组织大型商务数据库和支持商务智能的技术。OLAP数据库分为一个或多个多维数据集，每个多维数据集都由多维数据集管理员组织和设计，以适应用户检索和分析数据的方式，从而更易于创建和使用所需的数据透视表和数据透视图。OLAP的应用主要是针对用户当前及历史数据进行分析，辅助商业决策。其典型的应用有对银行信用卡风险的分析与预测、公司精准策略的制定等。其优势是能够进行大量的查询操作，对时间的要求不太严格。在数据仓库应用中，OLAP通常是数据仓库应用的前端工具，同时，OLAP工具还可以同计算分析工具和数据挖掘工具配合使用，以增强决策分析功能。

1.2.4　商业智能的研究内容和发展方向

1. 商业智能的研究内容

商业智能是以计算机高级技术为技术支撑、以现代管理技术为指导的应用型系统[①]，其研究热点主要包括体系结构、支撑技术以及应用系统三个方面。

（1）体系结构。它是指通过识别和理解数据在系统中的流动过程和数据在企业的应用过程中提供的商业智能系统的主框架。商业智能的体系结构包括：数据的预处理、数据仓库、数据分析和数据可视化等几部分。针对指定的应用会有相应的体系结构，从而使商业智能具有良好的性能。

（2）支撑技术。商业智能是一个20世纪90年代末期出现的跨学科的新兴领域，它的发展借助于两方面的先进成果：一是计算机技术，比如数据仓库技术、数据挖掘技术、联机分析处理技术、数据可视化技术、数据预处理技术和计算机网络技术等；另一个是现代管理技术，比如预测和统计等运筹学方法，供应链管理、企业资源计划、客户管理等管理理论与方法，此外，还有目前在研究领域比较热的建模技术与方法。支撑技术的研究主要围绕两部分展开：决策支持工具研究和企业建模方法研究。其中决策分析工具的研究还包括对各种分析方法的研究，而企业建模则是解决如何建立特定企业模式的辅助工具。

（3）应用系统。应用系统主要是当问题出现后，要根据提出的解决方案或方式决定具体的解决方法以及商业智能系统需要具备的功能，其研究重点主要是分析各个应用领域所面临的决策性问题。IT技术、人工智能等技术的不断发展，为商业智能的完善提供了强大的技术支持。

当前，商业智能在企业运营的相关领域及其他很多领域形成了其特有的体系，并且应用广泛。其中具有代表性的有：人力资源管理（HRM）、企业资源计划、企业性能管理（BPM）、客户关系管理、电子商务（E-Business）以及供应链管理。

2. 商业智能的发展趋势

（1）注重人性化，逐渐“傻瓜”化

今后商业智能的门户将更加注重人性化，功能也会逐渐“傻瓜”化，强调易用性，更加开放以及稳定性更高。此外更加重视整合众多信息来源，使人与人之间的沟通与合作更加便捷，帮助可拓展的管理支撑平台框架进一步完善，从而实现从“人去找系统”到“系统找人”的全新理念的转变。

未来商业智能系统能帮助人们充分发掘和释放潜能，帮助合适的角色在合适的时间、地点里获得合适的知识和数据，并且帮助企业将数据和信息转变为一种意念、能力，从而指导人的行为。在这里“人性化”也可以称为是一种“自动化”，使管理系统的价值与作用得以最大地体现。

（2）不断集成，演变成门户化

与决策支持系统（DSS）相比，商业智能具有更加美好的发展前景。未来的商业智能将会全面集成信息服务，可以通过类似“门户”的技术对各项业务进行系统地整合，BI可以融

① www.ciotimes.com

合集成CRM、ERP与SCM等应用系统。同时也可以联结企业所有信息资源和信息系统以及工作人员，从而真正实现跨平台，最后演变成门户化。

（3）移动BI将成为新战场

工信部的统计数据显示，截止到2014年5月底中国的手机用户数量已达到12.56亿人，相当于中国90.8%的人都在使用手机。而在所有使用手机的人中，使用手机上网的用户数量为8.57亿人，占总数量的68.24%。这些数字还在不断地增长，可以大胆地预测，未来利用新技术，移动协同应用将成为BI新的增长点。何谓移动协同应用，即用户可以在智能手机移动平台上提交数据，并且获取分析报告，实现商务智能与数据分析无处不在、无时不在的实时动态管理。该技术将会给传统的BI带来巨大的改变。因此，移动BI协同应用将是未来管理的巨大亮点。当前国内一些领先的、主流的BI软件企业正在积极地利用现代手机移动技术，想必未来的BI移动办公以及无线掌控将使管理者可脱离时间地点的控制，随时随地、随心所欲地进行管理。

（4）结合云计算，在云中部署BI

近年来，云计算的发展如火如荼。由于云计算拥有极其强大的功能，因此，商业智能部署的主流方向将会是以云为基础的商业智能在线服务。

可以说各个BI厂商未来的生存线与云计算的发展息息相关。从另一个角度考虑，BI软件的发展以及受欢迎程度必须要使产品基于面向云规模架构的设计，并符合云运营模式。即使BI在向云迁移的这个过程中会遇到许多困难与挑战，“在云中部署BI”也不再是天方夜谭，越来越多的企业会将其业务应用置于云端。目前，BI专业厂商Informatica发布的Informatica BI数据集成平台也已经能同时部署在“云”网络或预装在系统中，为企业用户提供云端集成服务。据了解，该公司已经尝试向用户交付云服务。

1.2.5 商业智能与大数据的关系

大数据与商业智能既相互区别又相互联系[①]，大数据是商业智能概念外延的扩展、手段的扩充，而不是取代的关系，也不是互斥的关系。

1. 联系

二者目的相同。无论是大数据还是商业智能，其目的都是为分析服务，对数据进行全面的整合，发现新的商业机会。对用户来说，其目的主要是得到一份完整的解决方案，形成一个全面、完整的数据价值发现平台。该解决方案不仅要能对企业内部的业务数据进行收集、处理和分析，要能引入互联网上的非结构化数据，如浏览的信息、微博和微信等，还要将移动设备的位置信息利用起来。

大数据促进商业智能的发展。大数据能够对内、对外产生价值，同时在保护隐私、保护数据安全的情况下，能够在不同组织间自由流动，形成整个社会的数据基础设施，进而形成一个平台。传统的商业智能无法处理日益复杂的数据，因此对传统的商业智能的扩展将成为未来的焦点。

① http://www.kmcenter.org/html/s55/201309/03-14901_2.html.

目前，尽管大多数业内人士认为，在数据分析市场中大数据与商业智能就像两个行驶在不同轨道上的列车，在并肩前行的同时偶有交叉。但是在技术上，为提升数据分析能力、提高洞察力，大数据与商业智能之间早就开始了交流与互动，以期更好地发展成长。

2. 区别

商业智能与大数据服务的领域侧重点不一样。现在可大致将数据资产划分为三种类型，即企业内部的业务数据、公共服务机构的数据（如物联网相关数据）以及与互联网相关的数据（如网络日志、微博等）。商业智能的服务领域侧重于前两者，而大数据则侧重于第三者，即与互联网相关的数据处理与分析。

二者应用的难易程度不一样。很长时间以来，商业智能被认为是大企业专有，因此被称为是“贵族”。其最经典的架构是以数据仓库为基础，搭建使用专用设备的数据仓库，利用ETL工具进行数据的抽取、转换与建模，然后通过报表、驾驶舱等形式进行结果展示，在整个过程中，每个环节相比较于其他非商业智能形式都更加耗时耗资。相对而言，大数据采用通用硬件设备加开源软件就可以实现，成本低，主要面向互联网企业，因此被称为“草根”。

二者效率、可靠性和安全性发展程度不一样。众所周知，在过去的20多年间，传统商业智能从磁盘数据库转向内存数据库，从行式存储数据库转为列式存储数据库，数据仓库实施从延时多维变为实时抽取，软件架构也从对称多处理机（SMP）转为大规模并行处理系统（MPP）等。用户对数据处理和分析需求的不断增加推动了这些转变的发生。一个成熟的数据处理平台所必须具备的条件包括效率、可靠性和安全性，而在这些方面传统BI已经走过了近20年的发展历程，可以提供很多值得借鉴的技术和方法。因此，对于大数据而言，这是它首先需要学习和发展的地方。

二者生态系统的完善程度不一样。商业智能已经存在了20多年，相对而言更加成熟，其生态系统也相对更完善。此外，大量企业在传统的商业智能方面投入很多，很多业务都是围绕该系统进行的，在数据采集、处理、存储、分析以及可视化软件等方面开展了很多工作。相比较而言，大数据出现的时间较短，应用不及传统商业智能广泛，其生态系统也不如传统商业智能完善[①]。

1.2.6　商业智能与大数据的结合应用

随着大数据时代的到来，商业智能与大数据的结合越来越紧密，并且已经应用到了各行各业中，如图1.5所示。

商业智能与大数据相结合，在各行各业中得到了广泛应用，其典型应用主要体现在四个方面。

1. 产品销售管理（Product Sales Management）

产品销售管理包括产品销售影响因素分析、销售量分析、销售策略及产品销售方案的预测四个方面。首先，为方便分析产生了不同结果的销售模型的销售量及销售策略，对影响销售的因素进行分析和评估，根据不同的销售环境，对相应的产品销售方案帮助制定产品上架

① http://www.kmcenter.org/html/s55/201309/03-14901_2.html.

和下架计划，使企业营销额得以提高。可根据系统储存的产品销售信息建立总体销售模型和区域、部门销售模型。除此之外，还可以通过对历史数据分析，建立预测模型，提高销售量。

图1.5　商业智能与大数据结合的应用

2. 事实管理（Management by Fact）

无论是目标管理还是例外管理，都需要用事实说话，用事实予以支持。过去，在信息缺乏的年代，管理层更多地是依靠个人的经验和直觉进行管理以及制定决策。而在当今知识经济时代，在每天的交易之中，维持企业营运的ERP系统已积累了庞大的事实与知识，这时就需要进一步对这些事实与知识充分分析并利用，结合企业目标、例外与事实，查询并探测相关信息，以便更好地决策。这些商业智能系统就能做到。

因此，企业必须实施事实管理，不靠个人经验和直觉，以了解企业每日的商务情况信息为基础，借助于商业智能进行科学决策。

3. 异常处理（Management by Exception）

在实际运行中，总会有一些偏差产生，商业智能系统可以监测实际与计划目标的偏差，实时并持续地计算各种绩效目标，这是商业智能数据挖掘应用的典型案例。在出现偏差过大的情况时，系统会采取各种通讯方式在第一时间通知企业责任主管，帮助企业主管及时知晓偏差状况，降低企业风险，进而提高企业收益。其具体应用包括银行及保险等行业的欺诈监测、信用卡分析等。

4. 客户关系管理（Customer Relationship Management）

众所周知，顾客是企业生存发展的关键因素，客户关系管理自然就成为企业一项重要的工作。为了采取相应对策保持顾客数量，培养忠实顾客，维持良好的客户关系，企业可以通过商业智能的客户关系管理子系统对顾客消费习惯和消费倾向进行分析，以便提高顾客满意度。

1.3　大数据相关技术与应用概况

1.3.1　大数据的相关技术

大数据技术，就是从各种类型的数据中快速获取有价值信息的技术。大数据领域已经涌现出了大量新的技术，它们成为大数据采集、存储、处理和呈现的有力武器。大数据处理相关的技术一般包括：大数据采集、大数据准备、大数据存储、大数据分析与挖掘以及大数据展示与可视化等，如图1.6所示。

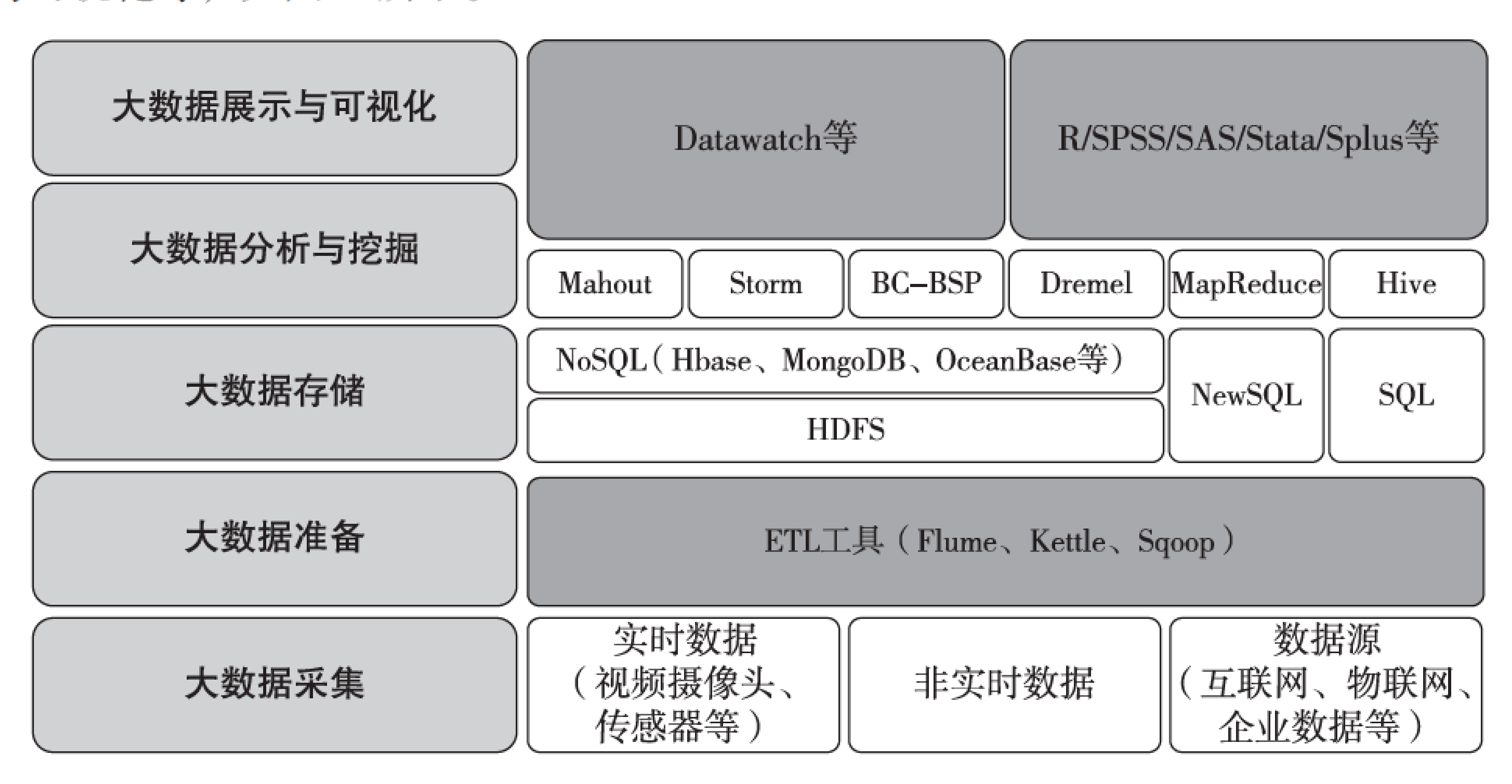


图1.6　大数据的技术体系

1. 大数据采集

大数据采集是指通过RFID射频数据、传感器数据、视频摄像头的实时数据、来自历史视频的非实时数据，以及社交网络交互数据及移动互联网数据等方式获得的各种类型的结构化、半结构化（或称弱结构化）及非结构化的海量数据。大数据采集是大数据知识服务体系的根本。大数据采集一般分为大数据智能感知层和基础支撑层。大数据智能感知层：主要包括数据传感体系、网络通信体系、传感适配体系、智能识别体系及软硬件资源接入系统，实现对结构化、半结构化和非结构化的海量数据的智能化识别、定位、跟踪、接入、传输、信号转换、监控、初步处理和管理等，需要着重攻克针对大数据源的智能识别、感知、适配、传输、接入等技术。基础支撑层：提供大数据服务平台所需的虚拟服务器，结构化、半结构化及非结构化数据的数据库以及物联网络资源等基础支撑环境，需要重点攻克分布式虚拟存储技术，大数据获取、存储、组织、分析和决策操作的可视化接口技术，大数据的网络传输与压缩技术，大数据隐私保护技术等。大数据采集方法主要包括：系统日志采集、网络数据采集、数据库采集和其他数据采集四种。

2. 大数据准备

大数据准备主要是完成对数据的抽取、转换和加载等操作。因获取的数据可能具有多种结构和类型，数据抽取过程可以帮助用户将这些复杂的数据转化为单一的或者便于处理的结

构，以达到快速分析处理的目的。目前主要的ETL工具是Flume和Kettle。Flume是Cloudera提供的一个高可用、高可靠、分布式的海量日志采集、聚合和传输系统；Kettle是一款国外开源的ETL工具，由纯Java编写，可以在Windows、Linux和UNIX上运行，数据抽取高效且稳定。

3. 大数据存储

大数据对存储管理技术的挑战主要在于扩展性。首先是容量上的扩展，要求底层存储架构和文件系统以低成本方式及时、按需扩展存储空间。其次是数据格式可扩展，满足各种非结构化数据的管理需求。传统的关系型数据库管理系统（RDBMS）为了满足强一致性的要求，影响了并发性能的发挥，而采用结构化数据表的存储方式，对非结构化数据进行管理时又缺乏灵活性。目前，主要的大数据组织存储工具包括：HDFS，它是一个分布式文件系统，是Hadoop体系中数据存储管理的基础；NoSQL，泛指非关系型的数据库，可以处理超大量的数据；NewSQL是对各种新的可扩展/高性能数据库的简称，这类数据库不仅具有NoSQL对海量数据的存储管理能力，还保持了传统数据库支持ACID和SQL等特性；HBase是一个针对结构化数据的可伸缩、高可靠、高性能、分布式和面向列的动态模式数据库；OceanBase是一个支持海量数据的高性能分布式数据库系统，实现了在数千亿条记录、数百TB数据上的跨行跨表事务。此外还有MongoDB等组织存储技术。

4. 大数据分析与挖掘

大数据分析与挖掘技术是基于商业目的，有目的的进行收集、整理、加工和分析数据，提炼有价信息的一个过程。数据分析是指通过分析手段、方法和技巧对准备好的数据进行探索、分析，从中发现因果关系、内部联系和业务规律，为商业目标提供决策参考。目前主要的大数据计算与分析软件包括：Datawatch，是一款用于实时数据处理、数据可视化和大数据分析的软件；Stata是一套提供其使用者进行数据分析、数据管理以及绘制专业图表的完整及整合性统计软件；Matlab是一款商业数学软件，一种用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境；SPSS“统计产品与服务解决方案”软件，为IBM公司推出的一系列用于统计分析运算、数据挖掘、预测分析和决策支持任务的软件产品及相关服务的总称；SAS是一个功能强大的数据库整合平台，可进行数据库集成、序列查询和序列处理等工作；Storm是一个分布式的、容错的实时计算系统；Hive是建立在Hadoop基础上的数据仓库架构，它为数据仓库的管理提供了许多功能，包括数据ETL（抽取、转换和加载）工具、数据存储管理和对大型数据集的查询和分析能力。此外还有R、BC-BSP、Dremel等计算和分析工具。

数据挖掘就是从大量的、不完全的、有噪声的、模糊的和随机的由实际应用产生的数据中，提取隐含在其中的，但又是潜在有用的信息和知识的过程。目前主要的数据挖掘工具有：Mahout，一个用于机器学习和数据挖掘的分布式框架，区别于其他的开源数据挖掘软件，它是基于Hadoop之上的；R是属于GNU系统的一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具。此外Datawatch、MATLAB、SPSS、SAS和Stata等都有着强大的数据挖掘功能。其中Datawatch桌面允许用户访问、抽取任何数据信息并将其转换为实时数据，以便显示、分析并与其他用户以及系统分享。企业用户可以在Datawatch桌

面上打开报告或文件，即点即选，数据立即就能提取出来。Datawatch系统创建了可复用模型，定义了数据到行和列的转换。仅需一次点击动作，用户就能将最新的数据集显示于仪表板上，并开始可视化数据发掘工作。

5. 大数据展示与可视化

大数据可视化技术可以提供更为清晰直观的数据表现形式，将错综复杂的数据和数据之间的关系，通过图片、映射关系或表格，以简单、友好、易用的图形化、智能化的形式呈现给用户，供其分析使用。可视化是人们理解复杂现象，诠释复杂数据的重要手段和途径，可通过数据访问接口或商业智能门户实现，以直观的方式表达出来。可视化与可视化分析通过交互可视界面来进行分析、推理和决策，可从海量、动态、不确定甚至相互冲突的数据中整合信息，获取对复杂情景的更深层的理解，供人们检验已有预测，探索未知信息，同时提供快速、可检验、易理解的评估和更有效的交流手段。目前，Datawatch、MATLAB、SPSS、SAS、Stata等都有数据可视化功能，其中Datawatch是数据可视化方面最流行的软件之一。完整的可视化分析系统的一个基本要素是：具有处理大量多变量时间序列数据的能力。Datawatch Designer可以提供一系列专业化的数据可视化方案，包括地平线图、堆栈图以及线形图等，让历史数据分析更简单、更高效。该软件能够连接传统的列导向和行导向的关系型数据库，从而支持对大型数据集进行快速、有效的多维分析。Datawatch提供了卓越的时间序列分析能力，是全球投资银行、对冲基金、自营交易公司以及交易用户必不可少的法宝。

1.3.2　大数据的应用概况

大数据应用自然科学的知识来解决社会科学中的问题，在许多领域具有重要的应用。早期的大数据技术主要应用在大型互联网企业中，用于分析网站用户数据以及用户行为等。现在，传统企业、公用事业机构等有大量数据需要处理的组织和机构，也越来越多地使用大数据技术以便完成各种功能需求。除了常见的商业智能和企业营销外，大数据技术也开始较多地应用于社会科学领域，并在数据可视化、关联性分析、经济学和社会科学领域发挥重要的作用。大数据应用基本上呈现出互联网领先，其他行业积极效仿的态势，而各行业数据的共享开放已逐渐成为趋势。

1. 大数据在互联网中的应用

互联网企业在大数据应用中处于领先地位，并逐步深入到其他行业中。互联网企业开展大数据应用拥有得天独厚的优势。互联网拥有大量的数据和强大的技术平台，同时掌握大量用户行为数据，能够进行不同领域的纵深研究。如谷歌、Twitter、亚马逊、新浪、阿里巴巴等互联网企业已广泛开展定向广告、个性推荐等较成熟的大数据应用。在此基础上，2012年，谷歌发布了其大数据的跨界应用——无人驾驶汽车。依靠庞大的道路信息数据（每秒钟会采集超过750MB的数据），无人驾驶汽车能够智能地选择路径以及自动驾驶。国内互联网企业以阿里巴巴为代表，其在2012年7月就已推出了数据分享平台“聚石塔”，为淘宝、天猫等平台上的电商提供数据云服务，并扩展到金融领域和物流领域。阿里巴巴基于对用户交易行为的大数据分析，提供面向中小企业的信用贷款。据透露，截至到目前，阿里巴巴已经放贷

300多亿元，而坏账率仅为0.3%左右。2013年5月，阿里巴巴成立的“菜鸟”网络物流，也是基于大数据平台，利用大数据平台的分析，联手各大物流企业，来选择最高效的送达方式。

2. 大数据在企业中的应用

大数据的挖掘和应用成为未来的核心技术，将从多个方面创造价值。大数据的重心将从传输和存储过渡到数据的挖掘和应用，这将深刻地影响企业的商业模式。据麦肯锡预测，大数据应用每年可潜在地为美国医疗健康业和欧洲政府分别节省3000亿美元和1000亿欧元，利用个人位置信息潜在地可创造出6000亿美元的价值，因此，大数据的应用是具有远超万亿美元的大市场。

企业的决策方法多以事实为基础，大量使用数据分析来优化企业运营的各个环节和流程，通过基于数据分析的业务优化和重组，把业务流程和决策过程中具有的潜在价值挖掘出来，从而达到节约成本、战胜对手、在市场中求生存的目标。大数据在企业中的分析包括顾客分析、商品分析、供应链和效率分析以及其他关乎企业绩效方面的分析。比如，电信运营商运用大数据进行智能管理中，基于用户、业务及流量分级的多维管控机制，以及精准的客户分析及营销（如套餐适配、离网预警、广告精准投放等）。这些应用大多数电信运营商早已执行，例如中国电信、西班牙电信、中国移动等，都已开展城市人口流量模型等工作。此外，电信业通过审视自身的数据优势，服务公共社会的应用逐步展开，像智慧城市、利用位置和轨迹信息服务社会、为智慧城市提供海量数据预测服务等。

3. 大数据在政府中的应用

大数据另外一个重要的应用领域是社会或政府。今天的城市面临着人口、就业和环境等各方面问题，许多宏观数据也是大数据分析的重要应用范畴。美国等发达国家的政府部门在开展大数据应用方面起了重要的表率作用，例如：美国能源部、联合国防部等6个联邦政府部门或机构投资了2亿美元，以开展大数据的政府应用。美国国防部开展了与网络安全相关的若干大数据项目，进行情报搜集和分析。美国国家卫生研究院着手建立健康与疾病相关的数据集、基因组信息系统、公众健康分析系统以及老龄化电子图书数据库等医疗大数据系统。国际上，早在2009年，联合国就启动了全球脉搏项目，跟踪和监控全球各地区的社会经济数据，采用大数据技术进行分析处理，以便更加及时地对危机做出反应。日本2012年开始对大数据进行专项调查，并将调查结果发布在《信息通信白皮书》里。2013年，日本总务省对大数据的发展现状进一步深入开展宏观和微观层面的调查，针对大数据的生成、流通与存储环节进行宏观的定量研究。在我国，大数据尚未上升到战略高度，应用案例也较少，在宏观大数据研究方面亟待加强。

4. 大数据在其他领域中的应用

大数据不仅在互联网、企业、政府中得到了广泛的应用，随着大数据的发展，大数据在医疗与生命科学研究、能源和司法执法等领域都得到了广泛的应用并不断扩展。比如：一个基因组序列文件大小约为750MB，一个CT图像大约为150MB的数据，一个标准的病理图则接近5GB。2010年，国家公布的“十二五”规划中提出要重点建设国家级、省级和地市级三级卫生信息平台，以及建设电子病历和电子档案两个基础数据库等。此外，各级医院也将加大

在数据中心，IT外包等领域的投入。随着医疗信息数据的增长速度成几何倍数不断发展，医院的信息存储越来越重要，医疗信息中心也将从关注传统计算领域转移到更加注重存储领域上来。

从2013年开始，电力、石油等能源细分行业纷纷拉开了大数据开发应用的序幕。大数据技术强调的是从海量数据中快速有效地获取有价值信息的能力，如何从海量数据中高效地获取数据，有效地深加工并最终应用到商业决策中是能源企业涉足大数据的目的。利用大数据可对业务进行分析，加工成有用的数据，进而全面掌控企业业务。如国网信通公司在北京亦庄的数据中心中，就设有10200个传感器。这些传感器及时采集数据，并被存储到云盘进行分析和利用。在大数据时代背景下，创新司法统计信息的收集与管理模式，深化司法统计数据的开发利用，对于更好地服务于审判管理，在更高的起点上推动人民法院工作实现新的发展，具有重要意义。

1.4　大数据热点问题与发展趋势介绍

1.4.1　大数据的热点问题

目前，大数据时代已经到来，不管是在学术界还是在产业界，人们都希望通过对大数据热点问题的研究，充分认识和了解大数据将要面对的关键性挑战和具有的独特价值，以便更好地把握投入方向，这对学术界、产业界以及用户具有指导价值。

1. 数据科学与大数据的学科边界

迄今为止，什么是大数据，在产业界和学术界并没有形成一个公认的定义，对大数据的内涵与外延也缺乏清晰的说明。另外，大数据是否就意味着全数据，还有待进一步讨论与澄清。最后，还需要为动态、高维和复杂的大数据建立形式化、结构化的描述方法，进而在此基础上发展大数据处理技术。而后者关注的是数据界与物理界、人类社会之间的关联与差异，探讨是否存在独立于应用领域的数据科学。如果存在数据科学，其学科问题的分类体系又是什么？目前已有的共识是，大数据的复杂性主要来自数据之间的复杂联系。另外，新型学习理论与认知理论等也应当是数据科学的重要组成部分。

2. 数据计算的基本模式与范式

大数据的诸多突出特性使得传统的数据分析、数据挖掘和数据处理的方式方法都不再适用。因此，面对大数据，需要有数据密集型计算的基本模式和新型的计算范式，需要提出数据计算的效率评估方法以及研究数据计算复杂性等基本理论。由于数据体量太大，甚至有的数据本身就以分布式的形式存在，难以集中起来处理，因此对于大数据的计算需要从中心化的、自顶向下的模式转为去中心化的、自底向上、自组织的计算模式。另外，面对大数据将形成基于数据的智能，我们可能需要寻找类似“数据的体量+简单的逻辑”的方法去解决复杂问题。

3. 大数据特性与数据态

这一问题综合了三个问题，即大数据的关系维复杂性、大数据的空间维复杂性和大数据的时间维复杂性问题。大数据往往由大量数据源头产生，而且常包含图像、视频、音频、数据流、文本与网页等不同的数据格式，因此其模态是多种多样的。主要来源于多模态的大数据之间存在着错综复杂的关联关系，这种异质的关联关系有时还动态变化，互为因果，因此导致其关联模式也非常复杂。大数据的空间维问题主要关注人、机、物三维世界中大数据的产生、感知与采集，以及不同粒度下数据的传输、移动、存储与计算。另外，还需研究大数据在空间与密度的非均衡态对其分析与处理所带来的理论与技术的挑战。大数据的时间维问题则意图在时间维度上研究大数据的生命周期、状态与特征，探索大数据的流化分析、增量式的学习方法与在线推荐。

4. 大数据的数据变换与价值提炼

大数据的数据变换与价值提炼即“如何将大数据变小”与“如何进行大数据的价值提炼”。前者要在不改变数据基本属性的前提下对数据进行清洗，在尽量不损失价值的条件下减小数据规模。为此，需要研究大数据的抽样、去重、过滤、筛选、压缩、索引和提取元数据等数据变换方法，直接将大数据变小，这可以看作是大数据的“物理变化”。后者可看作是大数据的“化学反应”，对大数据的探索式考察与可视化将发挥作用，人机的交互分析可以将人的智慧融入这一过程，通过群体智慧、社会计算、认知计算对数据的价值进行发酵和提炼，实现从数据分析到数据价值判定和数据制造的价值飞跃。

5. 大数据的安全和隐私问题

只要有数据，就必然存在数据泄露、数据窃取等与安全、隐私有关的问题。目前，大数据在收集、存储以及使用过程中都面临着重大的风险和威胁，大数据需要遵守更多、更合理的规定，但是传统的数据保护方法无法满足这一要求。因此，针对大数据的安全与隐私保护问题，还有大量的困难挑战亟需得到解决。具体挑战包括：大数据计算伦理学、大数据规模的密码学、分布式编程框架中的安全计算、安全的数据存储和日志管理、基于隐私和商业利益保护的数据挖掘与分析、数据计算的可信任度、实施安全/合规监测、强制的访问控制和安全通信、多粒度访问控制以及数据来源和数据通道等。总体而言，当前国内外针对大数据安全和隐私保护问题的研究还比较少，根据我国的国情，只有通过技术手段与相关的政策法规相结合才能更好地解决此类问题。

6. 大数据对IT技术架构的挑战

不管是存储系统、传输系统还是计算系统，大数据都提出了很多非常苛刻的要求，况且大数据平台本身也将是技术高峰，现有的数据中心技术很难实现大数据所提出的技术需求。譬如，数据的增长远远超过了存储能力的增长，对此目前需要解决的关键问题就是设计出最合理的分层存储架构。分布式存储架构需要结合scale-up式和scale-out式的可扩展性，因此对整个IT架构进行革命性地重构势在必行。此外，大数据平台（包括计算平台、传输平台和存储平台等）是大数据技术链条中的瓶颈，特别是大数据的高速传输，需要革命性的新技术。

7. 大数据的应用及产业链

大数据的研究与应用不是单一化的，应该与领域知识相结合，特别是在开展大数据研究的初期，计算机领域的科技工作者一定要虚心请教各领域的科技人员，从而真正了解和熟悉各领域数据的特征。根据不同的应用需求和不同的领域环境，大数据的获取、分析与反馈的方式也不尽相同。为此，针对不同行业与领域业务需求，首先需要展开业务特征与数据特征的研究，进行大数据应用分类与技术需求分析，然后构建从需求分析与业务模型，到数据建模、数据采集和总结反馈，最后到数据分析的全生命周期应用模型。其实，不同的应用环境和应用目标代表了不同的价值导向，这对于大数据的价值密度有很大的影响。

在大数据产业链方面，随着大数据的不断发展，很多数据都不知道如何运用，于是大量数据服务公司产生了。我国已经形成了大数据的“生产与集聚层——组织与管理层——分析与发现层——应用与服务层”产业链。

8. 大数据的生态环境问题

大数据被喻为21世纪的“新石油”，它是一种宝贵的战略资源，因此对大数据的共享与管理无疑是其生态环境的重要部分。所有权是大数据共享与管理的基础，而所有权既是技术问题，也是法律问题。因此，数据也是拥有权益的，对它的权益需要进行具体认定并进行保护，进而在保护好多方利益的基础上解决数据共享问题。为此，可能会遇到很多的困难，比如人们对法律或信誉的顾虑，保护竞争力的需要，以及数据存储的位置和方式不利于数据的访问和传输等。此外，生态环境问题受到政治、经济、社会、法律、科学等因素的交叉影响，因为大数据将对国家治理模式、组织和业务流程、企业的决策、个人生活方式都将产生巨大的影响，因此这种影响模式值得深入研究。

1.4.2　大数据的发展趋势

1. 大数据从概念化走向价值化

一方面，大数据将向更多新领域扩张，也会出现更多数据驱动的商业模式，更具体点说，互联网金融等将会成为大数据应用的新的商业模式，特别是基于海量数据的信用体系和风险控制，一定会冒出来。另一方面，资本高度关注大数据领域，相关的融资、并购与IPO纷纷出现，因此大数据从概念走向价值化成为大数据发展趋势中的最大趋势。

2. 大数据安全与隐私越来越重要

大数据安全不容忽视，这是因为大数据更容易成为网络中的攻击目标；对存储的物理安全性要求也会越来越高；大数据分析技术更容易被黑客利用；大数据引起了更多不易被追踪和防范的犯罪手段。个人隐私的问题也更为严重。个人的隐私越来越多地融入各种大数据中，大数据拥有者掌控了越来越多人的越来越丰富的信息。同时，有偿的隐私保护服务会被大众所接受。

3. 大数据分析与可视化成为热点

大数据规模大，难理解，分析过程离不开可视化技术，可视化将贯穿于大数据分析与

结果展示的全过程，可视化已经成为很多领域研究的议题。有了大数据以后，大规模、多角度、多视角与多手段的数据可视化，还有实时处理分析和大数据的处理方法贯穿了整个数据分析和数据展示的过程。

4. 数据的商品化和数据共享的联盟化

数据共享联盟有望逐步壮大，成为产业、科研和学术界一个环环相扣的支撑环节和产业发展的核心环节。另外，由于数据变成资源，成为有价值的东西，数据私有化和独占问题就是客观存在的，成为关注的焦点。数据产权界定问题日益突出，在数据权属确定的情况下，数据商品化将成为必然选择。

5. 深度学习与大数据性能成为支撑性的技术

在大数据时代，依靠高性能计算的支持，深度学习将会成为大数据智能分析的核心技术之一。基于海量智能的技术成为发展的热点，它利用群体智能和众包计算支撑大数据分析和应用，依赖于对捕捉到的数据的分析来做判断和决策，这将成为将要兴起的下一个浪潮。以分布式计算来支撑大数据分析是必经之路。在很多大数据的应用场合，基于物理资源的分散式应用会有更多的应用场景。

6. 数据科学的兴起

数据科学作为一个与大数据相关的新兴学科出现，各种大数据分析系统各有所长，在不同类型分析查询下，表现出不同的性能差异，使大家对数据科学兴起有了更具体的认识。目前，许多研究机构、学术团体和高校都在进行对大数据的研究以及大数据方面的学科建设和实验室建设，使得大数据成为一门真正的数据科学。

7. 大数据产业成为一种战略性产业

早在2011年，全球知名咨询公司麦肯锡发布了《大数据：创新、竞争和生产力的下一个前沿领域》的报告，预示了大数据产业将会成为本世纪具有决定性的产业。发展大数据产业，利用大数据分析提高国家经济决策和社会服务能力，保障国家安全成为各国的重要战略。除大企业成为大数据最活跃的群体外，一些拥有大数据的政府部门也纷纷利用积累的数据，采用大数据技术进行分析，产生了突出的效果。

8. 大数据生态环境逐步完善

虽然大数据生态环境目前还没有完善到令人满意的程度，但是它正在逐步完善。一方面，开源逐步成为主流；另一方面，大数据、云计算、物联网相互交融，开展大数据教育、计算机类相关的教育活动等，其中大数据教育更多是对人才方面的教育。

9. 大数据处理架构的多样化模式并存

在大数据处理方面，Hadoop/MapReduce框架一统天下的模式已被打破，实时流计算、分布式内存计算、图计算框架等并存；在大数据存储与管理方面，大数据的4V特征放大了以前海量数据的存储与管理的挑战；在性能提升方面，内存价格不断降低，使内存计算将成为解决实时性大数据处理问题的主要手段。

1.5　练习

1. 大数据的特征是什么?
2. 大数据的潜在价值是什么?
3. 在大数据时代这个大背景下，大数据将面临什么挑战?
4. 商业智能是什么?
5. 商业智能和大数据是什么样的关系?
6. 大数据的技术架构体系是什么?
7. 大数据在未来有什么样的发展趋势?

参考文献

[1] 陶雪娇，胡晓峰，刘洋. 大数据研究综述[J]. 系统仿真学报，2013，8（25）：143.

[2] 吴勇毅. 值得关注的BI未来新趋势[N]. 信息与电脑，2010：32–33.

[3] 邹大斌. 大数据动了谁的奶酪[N]. 计算机世界，2013，31.

[4] 刘业政，胡剑. 商业智能的核心技术及体系结构研究[J]. 合肥工业大学学报，2004，27（8）：885.

[5] 于希国. Hitachi UCP融合基础架构解决方案[N]. 中国计算机报，2013，1.

[6] 马天蔚. 商业智能三趋势[J]. 每周电脑报，2004（32）：42.

[7] 伍永锋. 商业智能及其技术[D]. 贵州大学，2008：40.

[8] 余长慧，潘和平. 商业智能及其核心技术[J]. 计算机应用研究，2002，19（9）：14–16.

[9] 李苹. EIM：精准商务智能的基础[J]. 软件世界，2006（19）：55.

[10] 阮京文. 宏观视野的大数据价值探索之路[J]. 广告大观（综合版），2013（12）：16.

[11] 张引，陈敏，廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展，2013（S2）：216–233.

[12] 何宝宏，魏凯. 大数据技术发展趋势及应用的初步经验[J]. 金融电子化，2013（6）：31–34.

[13] 谭琳. 大数据技术初探[J]. 科技创新导报，2014（4）：48.

[14] 王妍，柴剑平. 大数据及相关技术解读[J]. 广播电视信息，2014（2）：18–21.

[15] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013（01）：146–169.

[16] 陈健，冀超君. 大数据的应用及其发展趋势[J]. 山西科技，2014，29（2）：95–96.

[17] 栗蔚，魏凯. 大数据的技术、应用和价值变革[J]. 电信网技术，2013（07）：6–10.

[18] 黄哲学. 面向大数据的商务智能技术及应用[R]. 中国科学院深圳先进研究院，2013.

[19] 黄宜华. 大数据研究的技术层面与主要研究内容[R]. 南京大学计算机科学技术系.

[20] 罗诗惠. 大数据的应用和发展探讨[J]. 计算机科学，2014（2）.

[21] 张文涛. 大数据时代[R]. 阿里研究中心.

[22] 周倚平. 2010年商业智能研究分析报告[R].

[23] Tankard, C. Big data security. Network Security, 2012(7): 5-8.

[24] Smith, M, et al. Big data privacy issues in public social media. IEEE, 2012: 1-6.

[25] Sawant, N &.S.H. Big Data Application Architecture Q&A: A Problem-Solution Approach, 2013.

[26] Oikawa Lucas, Alberto Ryuga. Big Data on the Internet of Things. An example for the E-health.

[27] Big data: The next frontierfor innovation, competition, and productivity, McKinsey Global Institute, 2011.

[28] Pete. Warden. Big Data Glossary[M].

[29] 赵刚. 大数据：技术与应用实践指南[M]. 北京：电子工业出版社，2013.

[30] 大数据白皮书[R]. 中国计算机学会（CCF），2013.

[31] 大数据发展趋势预测[R]. 中国计算机学会（CCF），2014.

[32] http://blog.sina.com.cn/s/blog_6cf8fdd90101l2gu.html.

第2章 数据组织存储技术

随着大数据时代的到来，系统中需要存储的数据越来越多，数据也呈现出越来越复杂的结构。如何对海量数据进行组织、存储变得尤为重要，其中存储技术是关键。本章主要从数据存储的相关概念、数据存储技术的研究现状、海量数据存储的关键技术以及数据仓库等方面对数据的组织与存储技术进行详细地介绍。

2.1　数据存储概述

大数据应用的爆发性增长，直接推动了存储、网络以及计算技术的发展。随着经济全球化的不断发展，国际性的大型企业正在不断涌现，来自全球各地数以万计的用户产生了数以万计的业务数据，这些数据需要存放在拥有数千台机器的大规模并行系统上。同时，大数据也出现在日常生活和科学研究等各个领域，数据的持续增长使人们不得不重新考虑数据的存储和管理，海量数据存储对大数据时代至关重要。

数据存储是数据流在加工过程中产生的临时文件或加工过程中需要查找的信息。数据以某种格式记录在计算机内部或外部存储介质上。数据存储的命名需要反映出信息特征的组成含义。数据流要表现出动态数据的特征，反映的是系统中流动的数据；数据存储要表现出静态数据的特征，反映的是系统中静止的数据。

2.1.1　数据存储介质

数据存储介质主要包括磁带、光盘、硬盘三大类，在这三种存储介质的基础上分别构成了磁带机、光盘库、磁盘阵列三种主要的存储设备。未来高速海量数据存储的重要发展趋势是采用固态存储和全息存储。磁带机以其廉价的优势因而应用广泛；光盘库适用于保存多媒体数据和联机检索，应用也越来越普遍；磁盘阵列具有较高的存取速度和数据可靠性特点，成为目前高速海量数据存储的主要方式。

磁盘阵列（RAID，Redundant Array of Independent Disks）是冗余的独立磁盘阵列的英文缩写。冗余的目的是为了补救错失、保证可靠性，独立是指阵列不在主机内而自成一个系统。一般将RAID分为不同的级别，主要包含RAID0～RAID50等数个规范，每个规范的侧重点各不相同。最常用的级别是RAID0～RAID6。磁盘阵列通过在多个磁盘上同时存储和读取数

据来大幅度提高存储系统的数据吞吐量。磁盘列阵是由很多价格便宜的磁盘组合成的一个容量巨大的磁盘组，利用单个磁盘提供数据所产生的加成效果来提升整个磁盘的系统效果。通过控制和管理阵列控制器，磁盘阵列系统能够将几个、几十个甚至几百个磁盘连接成一个大磁盘，使总容量高达几百甚至上千兆，并且其速率也可以达到单个磁盘驱动器的几倍、几十倍甚至上百倍。磁盘阵列技术还有一个特点就是安全。它通过数据校验来提供容错功能，从而提供了更高的安全性。大多数RAID模式中的相互校验/恢复的措施都较为完备，有些是直接相互的镜像备份，用户数据一旦发生损坏，利用备份的信息即可使损坏的数据得以恢复，从而使RAID系统的容错度、系统的稳定冗余性以及用户数据的安全性都有很大提高。

2.1.2 数据存储模式

数据存储的方式主要有DAS、NAS、SAN和IP存储。

1. DAS存储方式

DAS（Direct Attached Storage，直接连接存储）是利用连接电缆，将外置存储设备连接到一台主机上。主机与存储设备间有多种连接方式：SCSI，小型计算机系统接口；ATA，先进技术附加设备；SATA，串行ATA；FC，光纤通道。在实际应用中，SCSI方式使用最多，随着服务器CPU的处理能力越来越强，硬盘的存储空间越来越大，组成阵列的硬盘数量也越来越多，SCSI通道成了I/O瓶颈。传统SCSI所提供的存储服务有很多限制，其中最关键的是：与服务器连接距离有限；可连接的服务器数量有限； SCSI盘阵受固化的控制器限制以至无法进行在线扩容。在直接连接式存储中，由主机操作系统对文件和数据进行管理，因为主机结构中包括数据存储部分。操作系统在对磁盘数据的读写与维护管理过程中会占用主机资源（诸如CPU、系统I/O等）。可以看出，这种方式的优点是，磁盘读写带宽的利用率高，中间环节较少，购置成本较低；其缺点是，数据存储占用主机资源，使主机的性能受到较大影响，扩展能力有限，同时主机系统的软硬件故障对存储数据的访问也会直接造成影响。

2. NAS存储方式

NAS（Network Attached Storage，网络附加存储）是能够对不同应用服务器和主机进行访问的技术，是一种能够把独立且分布的数据整合为集中化管理的、大型的数据中心的技术。NAS采用独立于服务器，单独为网络数据存储而开发了一种文件服务器来连接所存储设备，自形成一个网络。这样，数据存储就不再是服务器的附属，而是作为独立的网络节点存在于网络中，为所有的网络用户共享。

NAS的存储方法是部件级的，可以作为网络的一个节点存在，借助于双绞网线直接连接到IP网络上。NAS没有地域限制，具有支持远程实时访问、备份、操作等特性，这让它更容易部署。此外，NAS还具有安装简单、容易扩展、方便维护、安全可靠、低成本等特点。

NAS通信是按照TCP/IP协议来进行的，采用业界标准文件共享协议（如：NFS、HTTP和CIFS）来实现共享，数据传输方式为文件。借助于NAS自带的文件管理系统，安装在不同操作系统（如APPLE系统、Windows系统、Linux或UNIX）的客户机可以使用同一文件管理系统，使得异构平台之间的数据共享得以真正地实现。因此NAS存储方式具有良好的异构平台

兼容性。

3. SAN存储方式

SAN（Storage Area Network，存储区域网络）是指通过专用高速网将一个或多个网络存储设备和服务器连接起来的专用存储系统，未来的信息存储将以SAN存储方式为主。Fibre Channel（FC，光纤信道）协议是它的核心技术，该技术的建立是为了解决传统SCSI传输的距离限制。SAN支持HIPPI、IP、IPI、ATM和SCSI等多种高级协议，是ANSI为信道I/O接口和网络建立的一个标准集成。为实现在同一个物理连接上传送多种协议，光纤信道协议将网络和设备的通讯协议与传输物理介质隔离开来，这也是光纤信道协议的最大特性。

完全采用光纤连接是SAN技术的一大特点，从而保证了巨大的数据传输带宽，目前其传输距离可达到100km，数据传输速度也达到了4Gbit/s。一条单一的FC环路最大可以承载126个设备。与传统技术相比，为使存储与服务器分开成为现实，SAN技术将存储设备从传统的以太网中隔离出来成为独立的存储局域网络，这也是SAN的最大特点。此外，在SAN中实现容量扩展、数据迁移、远程容灾数据备份功能都比较方便。采用SAN技术的存储设备性能高，提高了数据的可靠性和安全性，但是其设备的互操作性较差，构建、管理和维护成本高，而且只能提供存储空间共享而不能提供异构环境下的文件共享。

4. IP存储

IP存储（Storage over IP，SoIP）是一种替代光纤通道（FC）的基于以太网和IP存储网的技术，它使服务器可以通过IP网络连接SCSI设备，在IP网络中传输块级数据，就如同使用本地的设备一样，用户不用关心设备的地址或位置。网络连接方式主要是IP和以太网。

由于IP存储技术既有的成熟性和开放性，使企业在制定和实现“安全数据存储”的策略和方案时，有了更多的选择空间。IP存储的介入大大丰富了远程的数据备份、数据镜像和服务器集群等领域的内容。同时，在企业IT部门设计传统SAN方案时必须要面对的两个问题（产品兼容性和连续性），在IP存储中已经不存在了。更重要的是，基于IP存储技术的新型SAN，兼具传统SAN的高性能和传统NAS的数据共享优势，为新的数据应用方式提供了更加先进的平台结构。

IP存储技术主要有两个方面，即存储隧道和本地IP存储。下面简单地介绍一下这两个方面[①]。

存储隧道（Storage tunneling）技术为了解决两个SAN环境的互联问题，将IP协议作为连接两个异地光纤SAN的隧道，在传输过程中，光纤通道协议帧被包裹在IP数据包中，专用设备会解开传输到远端SAN后的数据包，将其还原成光纤通道协议帧。

存储隧道技术提供的是两个SAN之间点到点的连接通信，因为从功能上讲，这种技术与光纤的专用连接技术类似，所以，这种存储隧道技术也被称为黑光纤连接（Dark fiber optic links）。要实现这种技术需要花费较高的成本，其专用性较强，缺乏通用性，而且较大的延迟在一定程度上也影响了性能。不过，可以将现有的城域网和广域网充分利用是其最大的优

① http://news.watchstor.com/news-29518.htm.

势，而这一优势正好满足了宽带资源的需要，使其进一步地充分利用成为可能。但是，另一方面，虽然存储隧道技术是利用IP网络进行传输的，但无法充分利用IP网络管理和控制机制相对完善这一优势。这导致目录服务、流量监控、QoS等一些很好的管理控制机制无法应用到存储隧道这种技术中，其主要原因在于，IP网络智能管理工具无法识别嵌入到IP数据包中的光纤通道协议帧。所以，企业IT部门的系统维护人员，对包含存储隧道的网络环境进行单一界面的统一集中化管理的可能性很小。总体来说，存储隧道技术虽然借用了一些IP网络的成熟性优势，但还是要依赖昂贵而复杂的光纤通道产品。

本地IP存储（Native IP-based Storage）技术为使网络和存储无缝融合，在IP协议中直接集成了如SCSI和光纤通道等现有的存储协议。即在传统的SAN结构中，将光纤通道协议替换成IP协议，构建新型SAN系统IP-SAN，使其在技术上与LAN一致，而在结构上与LAN隔离，而不是在物理上可以在企业IT系统中，合成一个将存储网络和传统的LAN整合在一起的网络。

在这种新型的IP-SAN中，用户可以直接把以往用户在IP网络上获得的维护经验、技巧应用到IP-SAN上，不仅能够保证性能，又有效地降低了成本。借助于随处可见的IP网络工具，IP-SAN可以方便轻松地进行网络维护。此外，与光纤技术培训相比，维护人员的培训工作也简单快捷许多。

本地IP存储技术还具有非常明显的优势。一方面，在本地IP存储技术中，用户接触到的是诸如IP协议和以太网这样比较熟悉的技术内容，并且，各种IP通用设备使得用户的选择空间变得非常广。实际上，充分利用现有设备是本地IP存储技术的设计目标之一。因此，可以在IP-SAN中充分利用传统的SCSI存储设备和光纤存储设备。另一方面，本地IP存储技术所具备的一体化的管理界面，也可以完全整合IP-SAN与IP网络。

在IP-SAN中，只要是主机和存储系统都能提供标准接口，无论哪一位置的数据都可由任意位置的主机进行访问，不管是在相隔几米的同一机房中，还是在相距数千米外的异地，这使得本地存储和远程存储的界限更加模糊。此外，本地IP存储技术的访问方式既可以与NAS结构中的通过NFS、CIFS等共享协议访问类似，也可以与本地连接和传统SAN中的通过本地设备级访问类似[①]。

2.1.3 大数据存储存在的问题

随着结构化数据和非结构化数据数量的不断增长，以及分析数据来源的多样化，之前的存储系统设计已经无法满足大数据应用的需求。对于大数据的存储，存在以下几个不容忽视的问题。

1. 容量

大数据时代存在的第一个问题就是“大容量”。“大容量”通常指的是可达PB级的数据规模，因此，海量数据存储系统的扩展能力也要得到相应等级的提升，同时，其扩展还必须简便，为此，通过增加磁盘柜或模块来增加存储容量，这样可以不需要停机。在解决容量问题上，不得不提及LSI公司的全新Nytro智能化闪存解决方案，采用这种方案，客户可以将数据库

① http://news.watchstor.com/news-29518.htm.

事务处理性能提高30倍，并且具有超过每秒4GB的持续吞吐能力，非常适合于大数据分析。

2. 延迟

“大数据”的应用不可避免地存在着实时性问题，尤其是涉及到网上交易或金融类的应用。“大数据”应用环境通常像HPC（高性能计算）那样需要较高的IOPS性能。正如改变了传统IT环境一样，服务器虚拟化的普及也对高IOPS提出了需求。为了迎接这些挑战，各种模式的固态存储设备应运而生，小到简单的在服务器内用做高速缓存的产品，大到全固态介质可扩展存储系统。通过高性能闪存存储，自动、智能地对热点数据进行读/写，高速缓存的系列产品如LSI Nytro都在蓬勃发展。

3. 安全

像金融数据、医疗信息以及政府情报等这些特殊行业的应用，都有自己的安全标准和保密性需求。对IT管理者来说，这些都是必须遵从的。但是，在过去没有需要多类数据相互参考的情况，而现在大数据的分析往往需要对多种数据混合访问，这就催生出了一些新的、需要考虑安全性的问题。此处不得不提及利用DuraClass™技术的LSI SandForceR闪存处理器，它实现了企业级的闪存性能和可靠性，实现了简单、透明的应用加速，既安全又方便。

4. 成本

成本控制是正处于大数据环境下的企业的关键问题，只有让每一台设备都实现更高的“效率”，同时减少昂贵的部件，才能控制住成本。目前，进入主存储市场的重复数据删除、多数据类型处理等技术，都可为大数据存储应用带来更大的价值，提升存储效率。在数据量不断增长的环境中，通过减少后端存储的消耗（例如LSI推出的Syncro™MX-B机架服务器启动盘设备），可以为企业减少成本，即使只降低了几个百分点，这样的服务器也能够获得明显的投资回报。现今，数据中心使用的基于传统引导方式的驱动器不仅故障率高，而且具有较高的维修和更换成本。如果用Syncro™MX-B机架服务器取代数据中心的独立服务器引导驱动器，则其可靠性能提升高达100倍。并且由于对主机系统是透明的，它能为每一个附加服务器提供惟一的引导镜像，简化了系统管理，提升了可靠性，并且节电60%，做到了真正的成本节省。Hadoop通常以集群的方式运行在廉价服务器上，也可以有效控制海量数据处理和存储的成本

5. 数据的累积

大数据应用大都会涉及法规遵从问题，这些法规通常要求数据保存几年或者几十年。例如为了保证患者的生命安全，医疗信息通常会被保存不少于15年，财务信息通常需要保存7年。因为对数据的分析大都是基于时间段进行的，任何数据都是历史记录的一部分，所以，有些大数据的存储希望能被保存得更久一点。数据被保存的时间越长，数据积累的就越多。要实现数据的长期保存，就要求大数据存储系统具有能够持续进行数据一致性检测的功能，以及其他保证长期高可用的特性，同时还要实现直接在原位进行数据更新的功能。

6. 灵活性

通常大数据存储系统的基础设施规模都很大，为了保证存储系统的灵活性，使其能够随

着应用分析软件一起扩容及扩展，必须经过详细设计。由于在大数据存储环境中，数据会同时保存在多个部署站点上，因此，没有必要进行数据迁移。一个大型的数据存储基础设施必须能够适应各种不同的数据场景与应用类型，因为，它一旦开始投入使用，就难以进行调整了。

7. 应用感知

目前，有些大数据用户已经在开发一些针对应用的基础设施，例如针对政府项目开发的系统、大型互联网服务商创造的专用服务器等。应用感知技术在主流存储系统领域的应用越来越普遍，它是改善系统效率和性能的重要手段。因此，应用感知技术也应该在大数据存储环境中使用。

8. 针对小用户

在大数据环境下，不仅仅是一些特殊的大型用户群体依赖大数据，作为一种商业需求，在不久的将来，小型企业也一样会应用到大数据。因此，为了吸引那些对成本比较敏感的小用户，一些存储厂商已经在开发小型的大数据存储系统了。

2.2 数据存储技术研究现状

目前，传统关系型数据库仍然是大部分互联网应用数据存储管理的主要选择，对数据的分析处理则通过编写SQL语句或MPI程序来完成。在用户和数据规模都相对较小的情况下，传统数据库系统尚可以高效运行。但是，在用户数量、存储管理的数据量都不断增加的情况下，如何应对更大规模的数据和满足更高的访问量，是许多热门的互联网应用在扩展存储系统时都会遇到的问题。

2.2.1 传统关系型数据库

在数据存储管理发展史上，传统关系型数据库是一座重要的里程碑。随着互联网时代的到来，之前主要集中在金融、证券等商务领域的数据存储管理应用模式已经不适用了。金融、证券等商务领域的数据处理对数据查询分析能力的便捷性、按照严格规则处理事务能力的速度、多用户访问的并发性以及保证数据的安全性有较高要求。正是针对这些要求，传统关系型数据库的设计具有这样一些特点：数据组织形式结构化、一致性模型严格、查询语言简单便捷、数据分析能力强大以及程序与数据独立性较高。也正是由于这些优点，传统关系型数据库得到了广泛的应用。

随着互联网时代的到来，需要处理的数据已经远远超出了关系型数据库的管理范畴，各种非结构化数据（包括博客、标签、电子邮件等超文本以及图片、视频与音频等）逐渐成为需要存储和处理的海量数据的重要组成部分。互联网数据快速访问，大规模数据分析的需求在关系型数据库中已经不能得到满足。主要表现在以下几个方面。

1. 应用场景局限

互联网应用主要面向的是半结构化和非结构化数据，这类应用与传统的金融、经济等应

用不同，它们大多没有事务特性，不要求保证严格的一致性，这本身就与传统关系型数据库的设计初衷不相同。虽然传统的数据库厂商也根据海量数据应用的特点针对性地提出了一系列改进方案，但是传统数据库在应对互联网海量数据存储效果上并不理想，因为这些解决方案并没有真正地从互联网应用的角度去设计。

2. 快速访问海量数据的能力被束缚

关系模型是一种按内容访问的模型，它是关系数据库的基础。在传统的关系型数据库中，行的值由相应的列的值来定位。这种访问模型会影响快速访问的能力，因为在数据访问过程中引入了耗时的输入输出。传统的数据库系统为了提高数据处理能力，一般是通过分区技术（水平分区和垂直分区）来减少查询过程中数据输入输出的次数，从而缩短响应时间。但是，这种分区技术对海量数据规模下的性能改善效果却并不明显。

Web 2.0中的许多特性都与关系模式中的严格范式设计相矛盾。例如，标签的分类模型是一种复杂的多对多关系模型，如果按照传统关系数据库范式设计要求——消除冗余性，就要将标签和内容存储在不同的表中，这就会导致系统性能的低下，因为对标签的操作需要跨表完成（在分区的情况下，还可能需要跨磁盘、跨机器操作）。

3. 对非结构化数据的处理能力不足

传统的关系型数据库对非结构化数据（视频、网页等）的支持度较差，只局限于一些结构化数据中（数据、字符串等）。随着硬件技术的快速发展、互联网多媒体交流方式的广泛推广以及用户应用需求的不断提高，处理庞大的音频、视频、图像与邮件等复杂数据类型的需求日益增长，用户对这些数据的处理要求也不只满足于简单的存储，而上升为识别、检索以及深入加工，对于这类需求传统数据库早已显得力不从心了。

4. 扩展性能差

在海量规模下，扩展性差是传统数据库的一个致命弱点。一般通过向上扩展（Scale up）和向外扩展（Scale out）来解决数据库扩展的问题。这两种方式分别从两个不同的维度来解决数据库在海量数据下的压力问题。向上扩展是通过升级硬件来提升速度，从而缓解压力。向外扩展则是按照一定的规则将海量数据进行划分，再将原来集中存储的数据分散到不同的物理数据库服务器上。在向外扩展的理念指导下，分片（Sharding）成为传统数据库的一种解决扩展性的方法。通过叠加相对廉价的设备，分片在存储和计算方面进行了扩展，不再受单节点数据库服务器输入输出能力的限制，提高了快速访问能力及提供了更大的读写带宽。但是，这种解决扩展性的方案在互联网的应用场景下仍然存在着一定的局限性。例如，这会要求互联网应用实现复杂的负载自动平衡机制，因为数据存储在多个节点上时，就要考虑负载均衡的问题，从而会花费较高代价；由于数据库范式规定严格，数据被表示成关系模型，从而难以被划分到不同的分片中；而一些数据可用性和可靠性问题也同样存在。

2.2.2　新兴的数据存储系统

通过对上节描述的传统关系型数据库的局限性可以看出，传统的数据库已经不能满足互联网应用的需求了。在这种情况下，一些主要针对非结构化数据的管理系统开始出现。这些

系统为了保障系统的可用性和并发性，通常采用多副本的方式进行数据存储。为了在保证低延时的用户响应时间的同时维持副本之间的一致状态，采用较弱的一致性模型（如最终一致性模型），而且这些系统也都提供了良好的负载平衡策略和容错手段。

1. HDFS

Hadoop是一个开源分布式计算平台，属于Apache软件基金会旗下，其核心是HDFS（分布式文件系统）和MapReduce，为用户提供分布式基础架构的系统底层细节。HDFS（Hadoop Distributed File System）是由Hadoop实现的一个分布式文件系统。它允许用户将Hadoop部署到低廉的硬件上，形成分布式系统，具有高容错性和高伸缩性等优点。通过MapReduce分布式编程模型，在不了解分布式系统底层细节的情况下，用户也可以开发并行应用程序。因此，利用Hadoop用户在组织计算机资源时能够更加轻松，进而能够搭建分布式计算平台，充分利用集群的计算和存储能力，完成大规模数据的处理。

HDFS由一个名称节点（NameNode）和N个数据节点（DataNode）组成，每个节点都是一台普通的计算机。在使用方式上HDFS与单机文件系统非常相似，它可以创建、复制和删除文件，创建目录，查看文件的内容等。但HDFS底层把文件切割成了Block，然后在不同的DataNode上分散地存储着这些Block，与此同时，为达到容错容灾的目的，每个Block可将数据复制数份存储于不同的DataNode上。整个HDFS的核心是NameNode，它通过一些数据结构的维护来记录每一个文件被切割成了多少个Block，可以从哪些DataNode中获得这些Block，以及各个DataNode的状态等重要信息①。

HDFS的设计目标有以下几点。

（1）硬件故障检测及恢复。硬件故障是常态，而不是异常，硬件故障的检测和自动快速恢复可以说是HDFS最核心的目标。构成整个HDFS系统的组件数目是巨大的——数百台或数千台存储着数据文件的服务器，每一个服务器都很有可能出现故障，这意味着在HDFS里总有一些部件是失效的。

（2）流式的数据访问。运行在HDFS上的应用程序不是普通文件系统上的普通程序，而是能流式地访问数据集。HDFS适合批量处理，而不擅长与用户交互式地处理。所以较之数据访问的低延时问题，它更看重数据吞吐量。

（3）简化一致性模型。HDFS简化了数据一致性问题，并使高吞吐量的数据访问成为可能，这主要得益于大部分的HDFS程序操作文件仅需一次写入，多次读取，经过创建、写入、关闭之后的文件不需要再进行修改。

（4）海量数据支持。运行在HDFS上的应用程序大都具有很大的数据集。HDFS的典型文件大小一般都在GB字节至TB字节。HDFS不仅可以用来优化大文件存储并且能提供集中式的高数据传输带宽，还能够使单个集群支持成百上千个节点。通常独立的Hadoop文件系统就能够支持上千万个文件。

（5）通信协议。HDFS系统的所有通信协议都是以TCP/IP协议为基础的。当明确配置了端口的名称节点和客户端连接之后，我们称它和名称节点的协议为客户端协议（Client

① http://blog.acdn.net/broadview2006/article/details/8783394

Protocal），而名称节点和数据节点（DataNode）之间则采用数据节点协议（DataNode Protocal）。

（6）异构平台间的可移植性。HDFS在设计的时候就考虑到异构软硬件平台的可移植性，可以简单方便地实现平台间的迁移，这种特性使得HDFS适合作为大规模数据应用平台。

Hadoop这个分布式计算平台可以让用户轻松使用和架构，在Hadoop上用户可以轻松地开发和运行处理海量数据的应用程序。Hadoop的优点一目了然：高可靠性、高扩展性、高容错性和高效性。基于Hadoop的应用因其突出的优势已经层出不穷，特别是在互联网领域的应用中。在互联网的不断发展中，不断涌现了一些新的业务模式，而对Hadoop的应用也从互联网领域拓展到了电信、银行、电子商务与生物制药等领域。

本书将在第4章中对Hadoop进行更加详细地介绍。

2. NoSQL

NoSQL是泛指非关系型、分布式和不提供ACID的数据库，它不是单纯地反对关系型数据库，而是强调键值存储和文档数据库的优点。

如上节所述，由于传统关系型数据库存在着灵活性差、扩展性差与性能差等原因，它们在处理数据密集型应用方面显得无能为力。最近出现的一些存储系统转向采用不同的解决方案来满足扩展性方面的需求，舍弃了传统关系型数据库管理系统的设计思想。人们普遍把这些没有固定数据模式的，可以水平扩展的系统统称为NoSQL（有观点认为将其称作NoREL更恰当），这里的NoSQL指的是Not Only SQL，而不是No SQL，它们是对关系型SQL数据系统的补充，而不是与之对立。

NoSQL系统普遍采用了以下一些技术。

（1）简单数据模型。大多数NoSQL系统采用的是一种更加简单的数据模型。这与分布式数据库不同，在这种更加简单的数据模型中，每个记录都有惟一的键，并且外键和跨记录的关系并不被系统支持，只支持单记录级别的原子性。这种一次操作获取单个记录的约束使数据操作可以在单台机器中执行，由于没有分布式事务的开销，极大地增强了系统的可扩展性。

（2）弱一致性。NoSQL系统的一致性是通过复制应用数据来实现的。由于NoSQL系统广泛应用弱一致模型，如最终一致性和时间轴一致性，减少了因更新数据时副本要同步的开销。

（3）元数据和应用数据的分离。NoSQL数据管理系统需要对元数据和应用数据这两类数据进行维护。但是这两类数据的一致性要求并不一样，只有元数据一致且为实时的情况下，系统才能正常运行；对应用数据而言场合不同，对其一致性需求也不同。因此，NoSQL系统将这两类数据分开管理，就能达到可扩展性目的。在一些NoSQL系统中甚至并没有元数据，解决数据和节点的映射问题需要借助于其他方式。

NoSQL借助于上述技术能够很好地解决海量数据带来的挑战。

与关系型数据库相比，NoSQL数据存储管理系统主要有以下几个优势。

（1）更简便。NoSQL系统提供的功能较少，避免了不必要的复杂性，从而提高了性能。相比较而言，关系型数据库提供了强一致性和各种各样的特性，但许多特性的使用仅发生在某些特定的应用中，大部分得不到使用的特性使得系统更复杂。

（2）高吞吐量。与传统关系型数据库系统相比，一些NoSQL数据系统的吞吐量要高得多。

（3）低端硬件集群和高水平扩展能力。与关系型数据库集群方法不同的是，NoSQL数据系统是以使用低端硬件为设计理念的，能够不需付出很大代价就可进行水平扩展，因此可以为采用NoSQL数据系统的用户节省很多硬件方面的开销。

（4）避免了对象—关系映射。许多NoSQL系统能够存储数据对象，如此就规避了数据库中关系模型和程序中对象模型相互转化的昂贵代价。

NoSQL向人们提供了高效、廉价的数据管理方案。许多公司开始借鉴Google的Bigtable和Amzon的Dynamo的主要思想来建立自己的海量数据存储管理系统，而不再使用Oracle甚至MySQL。现在一些系统，如Cassandra被Facebook捐给了Apache软件基金会，开始变成开源项目了。

目前市场上主流的NoSQL数据存储工具有：Bigtable、Dynamo、HBase、MongoDB、CouchDB和Hypertable。此外还存在着一些其他的开源的NoSQL数据库，如Neo4j、Riak、Oracle Berkeley DB、Apache Cassandra与Memcached等。

本书将在第3章中对NoSQL进行详细介绍。

3. NewSQL

NewSQL是对各种可扩展/高性能数据库的简称，这类数据库在保持了传统数据库支持ACID和SQL等能力的同时，还具有NoSQL对海量数据的存储管理能力。人们普遍认为系统的性能是由ACID和支持SQL的特性制约的，其实不然，系统性能是由一些其他的机制如缓冲管理、锁机制或日志机制等影响的。因此，只需优化这些技术，在处理海量数据时，关系型数据库系统也能表现出良好的性能。

这类NewSQL系统虽然内部结构变化很大，但它们都有两个显著的共同点：都支持关系数据模型和都是用SQL作为其主要的接口。目前的NewSQL系统大致有三类：采用新的架构、利用高度优化的SQL存储引擎和提供透明分片的中间件层。

如今已经出现了许多NewSQL数据库。例如：Google Spanner、VoltDB、RethinkDB、Clustrix、TokuDB和MemSQL等。

当然，NewSQL与NoSQL也有交叉的地方。比如，可以将RethinkDB看作是NewSQL数据库中MySQL的存储引擎，亦可看作是NoSQL数据库中键/值存储的高速缓存系统。现在一些NewSQL提供商为没有固定模式的数据使用自己的数据库提供存储服务，同时一些NoSQL数据库也开始支持SQL查询和ACID事务特性。NewSQL既能够提供SQL数据库的质量保证，也能提供NoSQL数据库的可扩展性。VoltDB就是这样一个NewSQL数据库，其开发公司的CTO宣称，VoltDB使用NewSQL的方法处理事务的速度比传统数据库系统快45倍。可以把VoltDB扩展到39个机器上，在300个CPU内核中每分钟处理1600万事务，其所需的机器数比Hadoop集群要少很多。

2.3 海量数据存储的关键技术

为了满足数据、用户规模的不断增长的需求，自适应的数据划分方式以及良好的负载均衡策略对于构建一个TB级乃至PB级的数据存储系统来说是必不可少的。而且，也需要在保证

系统可靠性的同时权衡数据的可用性及一致性，用以满足互联网应用对高吞吐率、低延时的要求。

2.3.1　数据划分

在分布式环境中，进行数据存储必须跨越多个存储单元。影响系统性能、负载平衡以及扩展性的关键因素之一就是数据的划分。系统必须在用户请求到来时将请求进行合理分发，这样才能提供低延时的系统响应，克服系统性能的瓶颈。哈希映射和顺序分裂是目前海量数据管理系统进行数据划分主要采取的两种方式。为了适应数据的多样性和处理的灵活性，在现在的互联网应用中，数据通常以键/值对方式进行组织。哈希映射这种数据划分方式带来的性能收益往往依赖于哈希算法的优劣性，因为它是根据数据记录的键值进行哈希运算，然后再根据哈希值将数据记录映射至对应的存储单元中。顺序分裂的数据划分方式是渐进式的。根据键值排序将数据写到数据表中，当数据表大小达到阈值后即可进行分裂，然后将分裂得到的数据分配至不同的节点上继续提供服务。这样，根据键值新流入的数据就能自动找到相应的分片并插入到表中。

Cassandra和Dynamo对数据的划分是通过一致性哈希映射方式进行的。这种方式在为系统带来良好的扩展性的同时，通过在数据流入时均匀地映射数据到对应的存储单元中，能够最大限度地避免产生系统热点。

Bigtable对数据的划分采用了顺序分裂的方式。这种划分方式是渐进式的，能够有效地利用系统资源提供良好的扩展性。但是频繁插入某个键值范围可能会导致负载热点的产生。区别于哈希映射方式的是，顺序分裂的数据和存储节点并不是直接映射的，为了集中管理这种分裂和映射行为，在Bigtable中需要有一个主控节点。因此，主控节点的管理能力限制了整个系统的扩展性。

PNUTS的数据组织结合了这两种方式，它既提供顺序表的组织方式，又提供哈希表的组织形式，采用了顺序分裂的方式，按照键或键哈希值来划分顺序表或哈希表中的数据。简而言之，PNUTS哈希表中的数据按照键的哈希值来有序存放。这些系统虽然根据不同的数据模型（顺序表、哈希表、键/值对等）来对数据进行组织，但它们都按照这些数据组织的特性实现了可扩展的数据划分。所以海量数据存储系统设计首要解决的问题应当是基于应用数据的特性，合理地对数据划分策略进行敲定，从而达到高可扩展性。

2.3.2　数据一致性与可用性

在分布式环境下，数据一致性为数据操作的正确性做出保证，而数据可用性则是数据存储的基石。一般情况下，为了解决数据的可用性问题往往会采用副本冗余、记录日志等方式。然而副本冗余又会带来数据一致性的问题。在运用副本冗余方式的分布式系统中，数据一致性及系统性的矛盾往往难以调和，需要在严格的数据一致性和系统的性能（如响应时间等）之间进行折中。有时在互联网应用需求下，要牺牲严格的数据一致性来调和这种矛盾，即为了保证高效的系统响应而允许系统弱化一致性模型，同时采用异步复制的手段用以确保数据的可用性。

Dynamo、Bigtable和PNUTS系统的数据高可用性主要是通过副本冗余的方式来保证的。然而，它们的具体实现并不完全相同。Dynamo采用的是整个系统中无主从节点之分的非集中的管理方式，其副本的异步复制就是在整个哈希环上通过gossip机制进行通信来完成的。而Bigtable和PNUTS采用的是集中管理方式，其服务节点内存中的数据可用性均是利用日志的方式来保证的。但是在数据存储可用性方面，两者又有不同，与Bigtable依赖于底层分布式文件系统的副本机制不同，PNUTS的数据冗余存储主要是采用基于发行/订阅（pub/sub）通信机制的主从式异步复制方式来实现的：先将数据同步至主副本，接着再通过发行/订阅机制异步更新至所有副本。

如上所述，Dynamo和PNUTS是需要跨数据中心部署的，均采用异步复制的方式进行副本更新，为维护系统的高性能而在某种程度上牺牲一定的数据一致性。由此可知，数据一致性、可用性及系统性能的权衡考虑与应用特性和部署方式紧密相关。

2.3.3 负载均衡

在分布式环境下，如何进行高效数据管理的关键问题是负载均衡。负载均衡主要包含两个方面的内容：数据的均衡与访问压力均衡。如前所述，在分布式环境中，采用一定的划分策略，例如哈希、顺序分裂等，将数据进行划分并存储于不同的节点之上，再由不同的节点来对用户的访问请求进行处理。但是用户访问请求的分布规律具有无法预测性，这最终会导致数据存储分布及节点访问压力的不均衡。由于存在数据分布和访问负载不均衡的情况，整个系统的性能在持续的数据加载压力以及频繁的并发访问下将会下降。因此海量存储系统需要有一套良好的均衡机制来保证数据加载的高吞吐率、系统响应的低延时以及系统的稳定性。

虚拟节点是一种能够使访问压力达到均衡的技术，它能够采用虚拟化的手段来单元化节点的服务能力，根据访问压力大小将压力较小的虚拟节点映射至服务能力较弱的物理节点上，对压力较大的节点则映射至能力较强的节点上。这样在访问压力达到均衡的同时，数据也会达到均衡状态。为了使数据在均衡过程中数据迁移的开销达到最小，Dynamo采用了虚拟化技术，通过量化节点的存储能力，使虚拟后的存储节点能够相对均匀地分布到集群哈希环上，从而有效地避免了数据在均衡过程中导致的全环的数据移动。在非集中式系统中，可以由任意节点发起这些均衡操作，并由gossip通信机制和集群中另外的节点来协调完成。

Bigtable与像Dynamo这样的非集中式管理方式不同，对各个子表服务器（tablet server）上的访问负载状态是由主控节点（master）来监控的。同样的，子表的分裂和迁移是利用主控节点来调度管理的，从而将访问压力均匀地分散到各个子表服务器上。Bigtable的数据底层存储运用的是分布式文件系统，以一种巧妙的方式避免了数据均衡的问题，因为其访问压力均衡过程中并不涉及存储数据的迁移操作。相似的方式同样被PNUTS用来均衡访问压力。不一样的是，PUNTS的数据底层存储是本地的文件系统或数据库系统，它在实施子表（tablet）的分裂及迁移之际，需要对存储数据进行迁移[①]。

显而易见，有效的数据划分方式是一柄双刃剑，一方面它为系统扩展性提供基础，另

---

① http://wenku.baidu.com/view/d3c6e584d4d8d15abe234ecd.html.

一方面也给系统带来了负载均衡的问题。因此，海量存储系统面临着这样一个挑战：如何在通过虚拟化节点或表分裂等方式更改数据分布格局，在访问负载均衡的同时，要避免数据迁移，或者至少尽量降低数据迁移量。

2.3.4　容错机制

分布式系统的健壮性标志是容错性。保证系统的可用性和可靠性的关键问题就是节点的失效侦测和失效恢复。

1. 失效侦测

在像Dynamo和Cassandra这样的非集中式系统中，为了解每个节点的活动状态，各个节点之间需要定期进行交互，从而完成对失效节点的侦测。而在集中式系统中，整个分布式系统的节点状态信息需要由专门的节点（部件）来维护，失效节点是否存在需要通过"心跳"机制来侦测。如Bigtable的失效侦测主要是采用分布式锁服务chubby追踪主控节点的子表节点的服务状态来实现的。PNUTS中节点失效是否存在的判断则主要来自子表控制器（tablet controller）部件维护的活动节点路由信息[①]。

2. 失效恢复

为了确保系统的可靠性与可用性，需要有相应的失效恢复策略来完成对系统中侦测到的失效节点的恢复。在分布式系统中，存在两种节点失效的情况：临时失效（例如网络分区等）和永久失效（例如磁盘损坏、节点死机等）。在副本冗余存储的分布式系统中，失效通常会造成失效节点内存中数据的丢失，通常解决这类问题的方法是日志重做。而在不同的系统中具体的失效恢复策略又有不同的特点。

在此以Bigtable为例。Bigtable并不区分是临时失效还是永久失效。Bigtable依赖主控节点通过"心跳"机制来完成对失效的侦测，即在限定时间内如果主控节点无法通过"心跳"机制获得从节点的响应就认为该从节点已经失效。就算临时失效的节点有可能与主控节点重新建立连接，主控节点也会停止这些节点，因为这些节点上的服务早已被分配到其他的节点上了。由于服务的迁移并不涉及存储数据的移动，也就不会带来额外的系统开销，故究竟属于何种失效状态也就不存在区分的必要。这种共享存储方式依靠底层的分布文件系统，也对系统的失效恢复进行了简化。

在集中式系统中，主节点各种失效恢复方式的差异是由其主从节点的功能差异导致的。主节点的失效将是灾难性的，因为它维护的是系统元信息。在集中式系统中，为了防止主节点的失效，通常是利用节点备份（多机、双机备份）。然而Bigtable的集群节点的状态信息主要依靠chubby来管理，整个系统存储的元信息则使用子表服务器来加以管理，从而将主节点的管理功能弱化，降低了主节点失效而引起灾难的可能性，与此同时也减小了主节点恢复的复杂性。

在非集中数据存储系统中，如Dynamo，其哈希方式的数据划分策略，使得系统中各个节点在作为存储节点的同时也作为服务节点，服务迁移的同时伴随着海量的数据迁移。因此系

① http://wenku.baidu.com/view/d3c6e584d4d8d15abe234ecd.html

统需要极其认真地应付各种各样的失效状态，在失效恢复过程中应当努力避免由于大规模迁移存储数据而导致的系统花销。基于上述原因，临时失效和永久失效在Dynamo中会被区别对待。

由上可以看出，失效侦测技术的选择与集群管理方式是集中式还是非集中式有着紧密的关系，该选择一般相对固定，但是失效恢复策略的实现却因应用而有所不同。系统的设计者可基于应用特性，权衡系统性能与数据一致性、可用性等多个影响因素来选择更合适的失效恢复策略①。

2.3.5 虚拟存储技术

虚拟存储②就是将硬盘，RAID等多个存储介质模块按照一定的手段集中管理起来，在一个存储池（Storage Pool）中统一管理全部的存储模块。站在主机和工作站的角度来看，就是一个分区或是卷，而不是多个硬盘，更类似于一个超大容量（例如大于1TB）的硬盘。这种能够把多种、多个存储设备统一管理起来，为使用者提供大容量且高数据传输性能的存储系统，称之为虚拟存储。根据虚拟存储的拓扑结构可将之分为对称式和非对称式两种。对称式虚拟存储技术是指将虚拟存储控制设备和存储软件系统、交换设备集成为一个整体，内嵌于网络数据传输路径之中；非对称式虚拟存储技术是指虚拟存储控制设备独立于数据传输路径之外。根据虚拟化存储的实现原理也可分为数据块虚拟与虚拟文件系统两种方式。

虚拟存储系统的结构如图2.1所示。共享存储系统由三大部分组成，即运行于主机的存储管理软件、互联网络、磁盘阵列等网络存储设备。

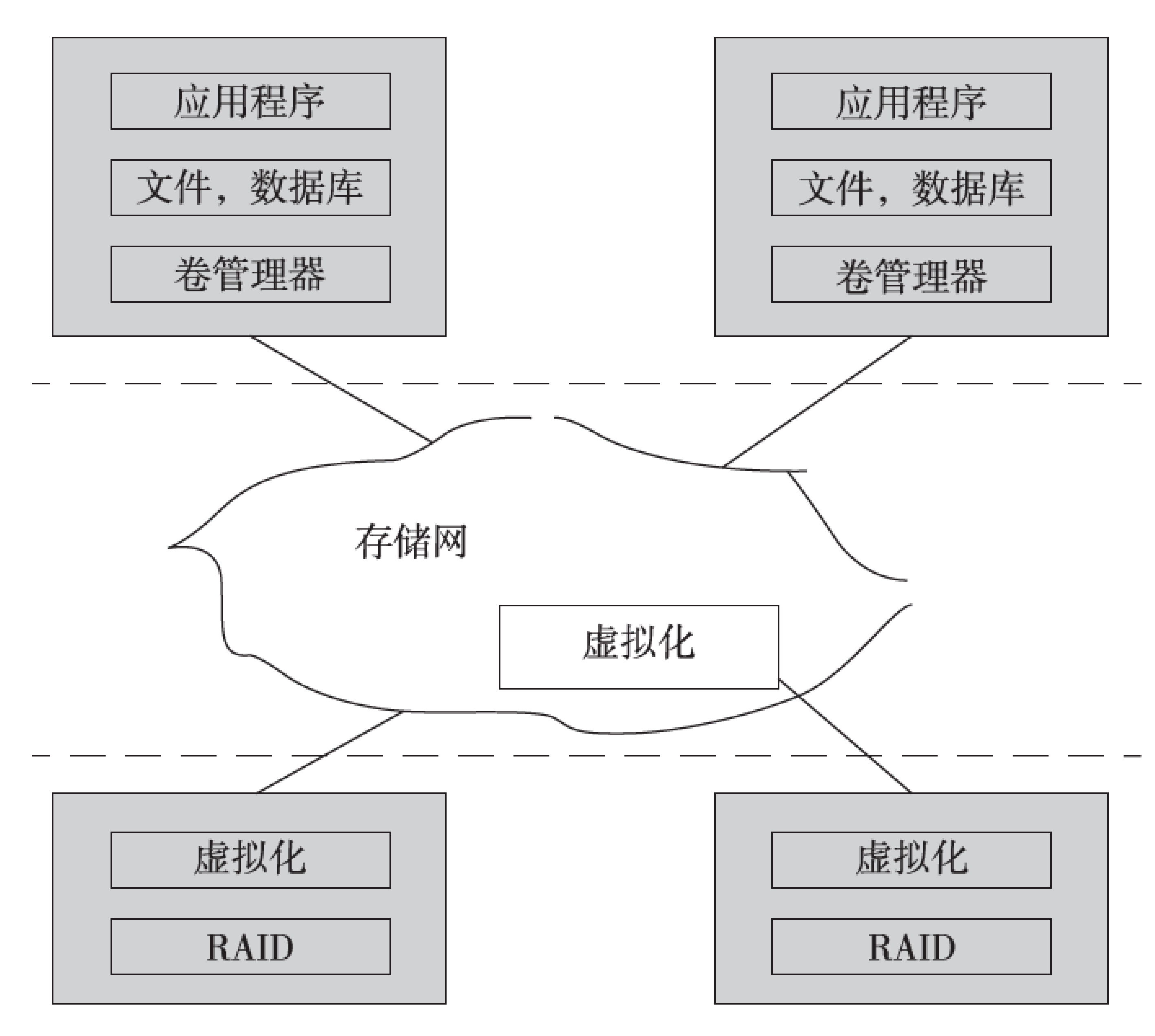


图2.1 虚拟存储系统的结构

与之对应，可以分别在共享存储系统的三个层次上实现存储虚拟化，即基于主机的虚

① http://www.docin.com/p-281269969.html.

② http://baike.baidu.com/view/2887659.htm?fr=aladdin.

拟存储、基于网络的虚拟存储和基于存储设备的虚拟存储[1]。各个层次的虚拟技术都各有特点，但其目的都是为了使共享存储更易于管理。

2.3.6　云存储技术

云存储是由云计算（cloud computing）概念延伸以及衍生发展而来的一个新的概念。云计算则是并行处理（Parallel Computing）、分布式处理（Distributed Computing）以及网格计算（Grid Computing）的发展，是借由网络把巨大的计算处理程序自动拆分为无数个相对较小的子程序，然后通过多部服务器所形成的庞大系统经运算分析之后，再把得到的处理结果传回给用户。借助云计算技术，网络服务提供者能够在短短的几秒内，处理数以千万计乃至上亿计的信息，达到提供与“超级计算机”一样强大的网络服务。与云计算的概念相类似，云存储是指凭借分布式文件系统、集群应用、网格技术等功能，通过应用软件将网络中大量不同类型的存储设备集合起来协同作用，实行共同对外提供数据存储以及业务访问功能的一个系统，这样既保证了数据的安全性，也节约了存储空间。简单说来，云存储就是将储存资源放到云上供用户存取的一种新兴方案，使用者不管处于何时何地都能够通过任何可连网的装置连接到云上，方便地存取数据。

云存储是一种新型态存储系统，它的产生是为了处理高速成长的数据量。云存储相比于传统的存储设备，不单单是一个硬件，更是一个由多个部分（如存储设备、网络设备、应用软件、接入网、公用访问接口、服务器、客户端程序等）组成的复杂系统。存储设备是各部分的核心，对外提供的数据存储以及业务访问服务主要通过应用软件来完成。云存储系统由4层组成：存储层、基础管理层、应用接口层和访问层，如图2.2所示。

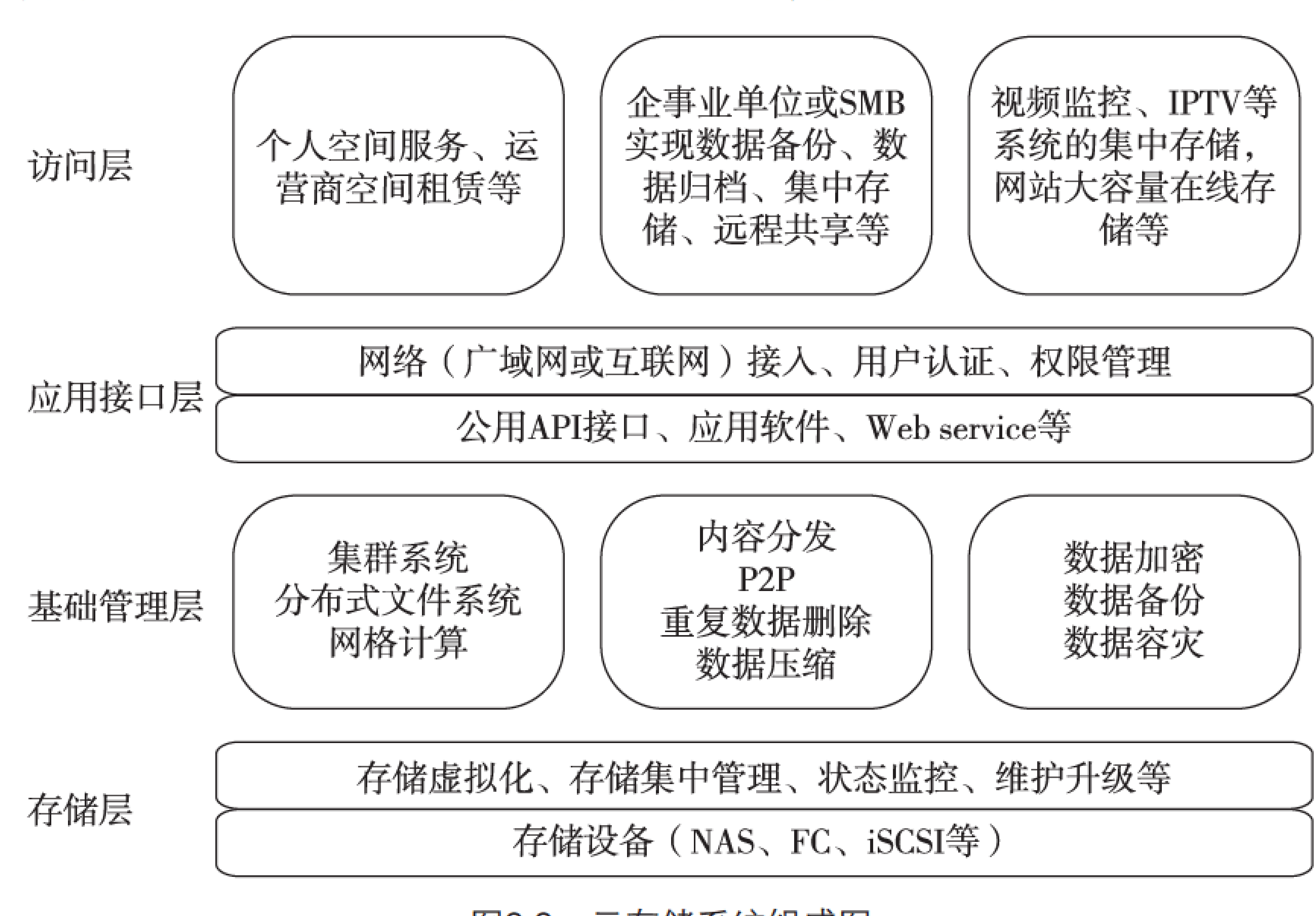


图2.2　云存储系统组成图

云存储的最基础部分就是存储层。存储设备可以是IP存储设备，如ANS和iSCSI等。也可

① http://baike.baidu.com/view/2887659.htm?fr=aladdin.

以是DAS存储设备，如SCSI或者SAS以及FC光纤通道存储设备等。云存储中的存储设备通常分布在不同地域且数量非常庞大，通过互联网、广域网或FC光纤通道网络把各个存储设备连接在一起。统一存储设备管理系统在存储设备的上一层，它能够完成多链路冗余管理，存储设备的逻辑虚拟化管理以及硬件设备的状态监控与故障维护。

云存储最核心、最难以实现的部分是基础管理层。基础管理层的主要功能是使云存储中多个存储设备之间可以协同工作，以便对外提供同一种服务，能够提供更大、更好、更强的数据访问性能，它所采用的技术主要有集群系统、分布式文件系统和网格计算等。为了保证云存储中的数据不会被未授权的用户所访问，它还提供了CDN内容分发系统以及数据加密技术。同时，为了确保云存储中的数据不丢失以及云存储自身的安全和稳定，它还采取了各种数据备份、数据容灾技术和措施。

云存储中灵活性最好的部分是应用接口层。根据实际业务类型的不同，不同的云存储运营单位开发的应用服务接口及提供的应用服务也不一样。例如在线音乐播放应用平台、网络硬盘应用平台、IPTV和视频点播应用平台、远程教学应用平台等。

用户获得云存储系统的授权后，就可以通过标准的公用应用接口进行登录并享受云存储服务。云存储提供的访问类型和访问手段会根据云存储运营单位的不同而有所不同。

2.4 数据仓库

数据仓库是决策支持系统和联机分析应用数据源的结构化数据环境。本节主要从数据仓库的相关概念、体系结构、设计与实施等方面介绍数据仓库，同时阐述了数据的抽取、转换和装载以及联机分析处理等数据仓库技术。

2.4.1 数据仓库的相关概念

随着计算机技术的快速发展以及企业界新需求的不断提出，数据仓库技术的出现便水到渠成。尤其是在大数据时代背景下，传统的以单一的数据资源，即以数据库为中心的数据库技术已不能满足数据处理多样化的需要。当前的数据处理大致可以分为两种：操作型处理（事务型处理）和分析型处理（或信息型处理）。操作型处理主要是为企业的特定应用服务的，是对数据库联机的日常操作，如对一个或一组记录进行查询和修改，人们普遍关心的是系统的响应时间以及数据的完整性和安全性。分析型处理主要是为管理人员的决策分析服务的，例如DSS、EIS和多维分析等，这类服务经常要访问大量的历史数据。两者之间的巨大差异导致了操作型处理和分析型处理的必然分离。

1. 数据仓库定义

数据仓库（Data Warehouse）的概念来自于W・H・Inmon在1992年出版的《建立数据仓库》（Building the Data Warehouse）一书。数据仓库是以关系数据库、并行处理和分布式技术为基础的信息新技术。除此之外，业界对数据仓库的定义还有很多种。

W・H・Inmon对数据仓库的定义：数据仓库是一个面向主题的、集成的、非易失的且随

时间变化的数据集合，用来支持管理人员的决策。

SAS软件研究所的观点：数据仓库是一种管理技术，旨在通过通畅、合理、全面的信息管理，达到有效的决策支持。

从数据仓库的定义可以看出，与数据库为事务处理服务不同，数据仓库是明确为决策支持来服务的。

2. 数据仓库的特点

从数据仓库的定义不难看出，数据仓库有四个主要特点：面向主题的、集成的、数据不可更新的、数据是随时间不断变化的。

（1）数据仓库是面向主题的

面向主题是数据仓库中最主要的一个特点。数据仓库的数据是按照一定的主题域进行组织的，排除了对决策无用的数据，提供特定主题的简明视图，而传统数据库是面向应用进行数据组织的。主题是一个在较高层次上将企业信息系统中的数据进行综合、归类并分析利用的抽象概念，它是数据归类的标准。每一个主题在逻辑意义上是与企业中某一宏观分析领域所涉及的分析对象相对应的。所谓面向主题的数据组织是一种在较高层次上对分析对象的数据进行一个完整、一致的描述，能够刻画各个分析对象所涉及的企业的各项数据以及数据之间联系的数据组织方式。其中较高层次指的是按照主题进行数据组织的方式，比面向应用的数据组织方式具有更高级别的数据抽象。

目前，数据仓库仍是采用关系数据库技术来实现的，即它的数据最终也用关系来表现，这里特别强调的是主题与面向主题这两个概念的逻辑意义。

在此以一家大型网上书店为例，来说明面向主题与传统的面向应用两者之间数据组织方式的差别，以便读者更好地理解主题这一抽象概念。该网上书店按照业务建立起了采购、库存、销售以及人事管理子系统，并按照各自的业务处理的要求，建立了数据库模式。

- 采购子系统：
  - ◆ 采购单（采购单号，出版社号，总金额，日期，采购员）
  - ◆ 采购单细则（采购单号，书号，采购单价，数量，总价）
  - ◆ 书（书号，书名，作者，类别，出版社号）
  - ◆ 出版社（出版社号，出版社名，地址，电话）
- 库存管理子系统：
  - ◆ 领书单（领书单号，领书人，日期）
  - ◆ 领书单细则（领书单号，书号，数量）
  - ◆ 进书单（进书单号，采购单号，进书人，收书人，日期）
  - ◆ 库存（书号，库房号，库存量，日期）
  - ◆ 库房（库房号，仓库管理员，地点，库存商品描述）
- 销售子系统：
  - ◆ 会员（会员号，会员名，支付帐号，会员等级，邮箱，电话）
  - ◆ 销售（订单号，会员号，总金额，日期，收货人，收货地址，收货电话）
  - ◆ 销售细则（订单号，书号，数量，销售单价，总价）

- 人事管理子系统：
  - 员工（员工号，姓名，性别，年龄，文化程度，部门号，职位号）
  - 部门（部门号，部门名称，部门主管，电话）
  - 职位（职位号，职位名称，职责描述）

按照面向主题的方式，数据的组织分为两个步骤：抽取主题以及确定每个主题所包含的数据内容。按照分析的需求抽取主题，每个主题有着各自独立的逻辑内涵，对应着一个分析对象。仍以上面网上书店为例，通过概括各个分析领域的分析对象，综合后得到各个主题。它所应有的主题包括书籍、出版社、会员等，这三个主题具体应包含如下内容。

- 书籍：
  - 书籍固有信息包括书号，书名，作者，类别等。
  - 书籍采购信息包括书号，出版社号，总金额，采购单价，数量，采购日期等。
  - 书籍销售信息包括书号，会员号，销售单价，销售日期，销售量等。
  - 书籍库存信息包括书号，库房号，库存量，日期等。
- 出版社：
  - 出版社固有信息包括出版社号，出版社名，地址，电话等。
  - 供应书籍信息包括出版社号，书号，采购单价，采购日期，数量等。
- 会员：
  - 会员固有信息包括会员号，会员名，支付帐号，会员等级，邮箱，电话等。
  - 会员购书信息包括会员号，书号，销售单价，购买日期，购买量等。

现在以“书籍”这一主题为例，可以看到关于书籍的各种信息都已综合在“书籍”主题中了。它主要描述的内容包括两方面：一，包含了书籍的固有信息，如书号、书名、作者以及类别等等书籍的描述信息；二，“书籍”主题中也包含有书籍的流动信息，如描述了某书籍采购信息、书籍销售信息以及书籍库存信息等。与网上书店原有数据库的数据模式相比，可以看出以下两点：一方面，在从面向应用到面向主题的转变过程中，丢弃了原来不必要的、不适于分析的信息，如有关采购单信息、领书单等内容就不再出现在主题中；另一方面，在原有的数据库模式中，关于书籍的信息是分散在各个子系统中的，根本没有形成一个关于书籍的完整的一致性描述，如书籍的采购信息存在于采购子系统中，书籍的销售信息则存在于销售子系统中，书籍库存信息却又存在于库存管理子系统中。而面向主题的数据组织方式就是强调要形成关于书籍的一致的信息集合，以便在此基础上针对“书籍”这一分析对象进行分析处理。

不同的主题之间也会有重叠的部分，这些重叠的部分往往是前面所说的第二方面的内容，如 “书籍”主题的书籍采购信息同“出版社”主题的供应书籍信息都来自采购子系统，它们是相同的，这表现了“出版社”和“书籍”这两个主题之间的联系；“书籍”主题的书籍销售信息则同“会员”主题中的会员购书信息都来源于销售子系统，这表现的是“书籍”和“会员”之间的联系。有两点特别需要注意的地方：第一，主题之间的重叠并不是同一数据内容的重复物理存储，而是逻辑上的重叠；第二，主题之间并不是两两重叠的，如“出版社”和“会员”两个主题间一般是没有重叠内容的，这表明了“出版社”和“会员”之间是

不直接发生联系的，它们之间的间接联系是通过“书籍”这一主题来体现的。

不同企业的主题也会不同，例如，对一家制造企业来说，销售、发货和存货都是非常重要的主题，而对于一家零售商来说，在付款柜台处的销售才是非常重要的主题。

主题只是一个逻辑的概念，它依然是基于关系数据库来实现的。在具体现实中，可将一个主题划分为多个表。但是数据仓库中的数据已经经过了一定程度的综合，而不再是业务处理的流水帐。如书籍表中一条记录是某段时期内该本书采购、销售情况的总和。

（2）数据仓库的数据是集成的

数据仓库中的数据是抽取自原有的、分散的数据库中的数据。分析型数据与操作型数据之间存在着很大的差别：第一，数据仓库的每个主题所对应的原数据是分散在不同数据库中的，且是与不同的应用逻辑捆绑在一起的，不可避免地会出现许多重复和不一致的地方；第二，数据仓库中的综合数据不是对原有数据的简单复制，因此，数据仓库建设中最关键、最复杂的一步就是必须在数据进入数据仓库前，消除数据中不一致及错误的地方，对数据进行统一，以保证数据的质量。这要完成两项工作：统一原数据中所有矛盾之处，以及进行数据综合和计算。

（3）数据仓库的数据是不可更新的

数据仓库中的数据在通常情况下是不会进行修改操作的，它所涉及的主要数据操作是数据查询，供企业决策分析之用。数据仓库中的数据并非是简单的联机处理的数据，而是不同时点的数据库快照的集合，和基于这些快照来统计、综合以及重组的导出数据，反映的是一段相当长的时间内历史数据的内容。存放在数据仓库中的数据是之前数据库中通过联机处理的数据集成后输入进去的，它是有存储期限的，一旦超过这个存储期限，便从当前的数据仓库中将这些数据删去。数据仓库管理系统DWMS相比DBMS而言要简单得多，因为数据仓库中通常只对数据查询进行操作。在数据仓库的管理中，几乎可以将DBMS中许多像完整性保护、并发控制这样的技术难点省去，但是数据仓库对数据查询的要求却比DBMS要高得多。因为数据仓库所涉及的主要数据操作就是数据查询，且其查询数据量往往很大，因此就要求运用各种各样繁复的索引技术。另外，由于企业的高层管理者是数据仓库服务主要面向的客户，因而对数据查询界面的友好性以及数据表示提出了更高的要求。

（4）数据仓库数据是随时间不断变化的

对应用而言，数据仓库中的数据是不可更新的，但并不意味着数据进入到数据仓库以后就永远不变，而是数据仓库的用户在进行分析处理时没有进行数据更新操作。数据仓库的这一特征主要有以下三方面的表现。

首先，随着时间的改变，新的数据内容将被增加到数据仓库里。数据仓库系统将捕捉OLTP数据库中变化的数据，生成OLTP数据库快照，数据经过统一集成后不断地追加到数据仓库中；捕捉到的新的变化数据以新的数据库快照形式增加到数据仓库中，而非对先前的数据库快照进行修改，先前已经获取的数据库快照是不再变化的。

其次，数据仓库会根据时间的变化将旧的数据内容逐渐删除。数据仓库中的数据并非永远存在于数据仓库中，与数据库中的数据一样，它也有一定的存储期限，当数据超过该期限时，过期的数据将会被自动删去。与数据库不同的是，数据仓库内的数据时限要长得多。为了适应DSS进行趋势分析的需求，数据在数据仓库中一般需要保存较长时间（如5～10年），

而在操作型环境中，数据一般只保存60～90天。

最后，数据仓库中含有许多与时间相关的综合数据，随着时间的不断变化，这些数据都需要重新进行综合。如经常需要根据时间段来综合数据，或是间隔一段时间片就对数据展开抽样等。为此，为了标明数据的历史时期，数据仓库数据的码键都要包含时间项。

3. 数据集市

数据仓库的工作范围和成本通常是巨大的。信息技术部门必须对所有的用户站在全企业的角度对待任何一次决策分析，这样会以巨大的金钱与时间为代价，这是许多企业不愿意或者不能够承担的。为了应对这种情况，一种提供更紧密集成的、拥有完整图形接口并且以价格优势吸引人的工具——数据集市就应运而生了。

数据集市（Data Marts）是一种更小、更集中的，具有特定应用的，部门级的数据仓库，可以按业务的分类来组织数据集市。数据集市一般针对具有战略意义的应用或者是具体部门级的应用，包含的是有关该特定业务领域的数据，如人力资源、财务、销售、市场等。数据集市非常灵活，不同的数据集市可以分布在不同的物理平台上，也可以逻辑地分布于同一物理平台上。因此，数据集市可以独立地实施，企业人员也可以快速获取信息。由于数据集市的结构简单，即使当其数据增长时，管理也较容易。当数据集市中加入了越来越多的主题时，就应将这些数据集市加以集成，最终建立起一种结构，即构成企业级数据仓库的数据。因此，可以把数据仓库作为一组数据集市来实施，每次实施一个。但是，在实施之前，应该有全局的观点，先使不同的数据集市中的数据内容有统一的数据类型、字段长度、精度和语义，这样，就可以使数据集市在扩展后集成为全企业级的数据仓库了。如果采用这种方法来实施，数据集市就是整个数据仓库系统的逻辑子集。

数据集市的特性主要有：①规模小；②特定的应用；③面向部门；④由业务部门定义、设计、开发以及管理和维护；⑤快速实现；⑥价格低廉；⑦投资快速回收；⑧工具集紧密集成；⑨更详细的、预先存在的数据仓库的摘要子集；⑩可升级到完整的数据仓库。

4. 数据粒度与分割

数据仓库极为重要概念之一是粒度。粒度指的是数据仓库中数据单位所保存数据的细化或者综合程度的级别。数据的粒度是数据仓库设计的一个主要方面。它深刻影响着存放在数据仓库中的数据量的大小和数据仓库可以回答的查询类型。越小的粒度其数据的细节反映程度越高，综合程度越低，这样就能回答越多的查询种类。相反，越大的粒度其数据的细节反映程度越低，综合程度就越高，只能回答综合性的问题。粒度并不是越大或者越小就好，针对不同类型的问题，对粒度大小的要求也不同。如果数据粒度设计得不合理，就会造成对大量细节数据进行综合并计算答案，使得效率变得十分低下；或者有时需要细节数据时却又不能满足。所以，要在查询效率和回答细节问题能力之间做好平衡。

因此，多重粒度在数据仓库中是不可避免的，应根据查询的需求合理地设计数据的粒度。数据仓库主要是面向DSS分析的，只有极少数的查询涉及到细节，其他绝大部分查询都是基于一定程度的综合数据之上。故而为了大幅度提高绝大多数查询性能，应该把大粒度数据存储在快速设备（如磁盘）上，而把小粒度数据存储在低速设备（如磁带）上，这样即使

需要对细节进行查询也能够满足。

数据仓库中的另一重要概念是分割。它是指把数据分散到不同的物理单元，从而可以分别处理，来提高数据处理的效率。通常把数据分割之后的数据单元称为分片。在实际分析处理时，对于存在某种相关性的数据集合的分析是极其常见的，例如对某个地区、某个时间或时段，又或者特定业务范围的数据的分析等。毫无疑问，将这些具有某种相关性的数据组织在一起将会大大提高效率，因此，需要对数据进行分割。

数据分割使数据仓库的开发人员和用户有了更大的灵活性，可以按照实际情况来确定分割的标准。通常可以根据地域、日期或业务范围等指标来进行分割，也可以根据多个分割标准的组合来加以分割。一般分割标准都要包含日期项，这样就能使分割十分自然且均匀。小单元内的数据在分割之后，就会变得相对独立，加快处理速度。数据分割使数据的索引、重组、重构、恢复、监控和顺序扫描变得更简单。

5. 元数据

与传统数据库中的数据字典类似，元数据（metadata）是数据仓库的一部分不可或缺的重要数据。它是“关于数据的数据”，描述的是数据仓库中数据的结构、内容、码以及索引等。在数据仓库中，元数据有着比数据字典更加丰富和复杂的内容，主要有两种元数据：第一种元数据包含了所有原数据项名、属性以及它在数据仓库中的转换，它是为了从操作型环境向数据仓库环境转换而建立的；第二种元数据称为DSS元数据，是在数据仓库中用来在终端用户的多维商业模型以及前端工具间建立映射，一般是为了开发出更加先进的决策支持工具而创建的。

元数据在数据仓库的建造以及运行中的作用极为重要，它描述了数据仓库中的各个对象，是数据仓库的核心，遍及数据仓库的各个方面。数据仓库中的元数据的主要作用有：定义数据仓库中有什么；指明数据仓库中信息的内容及位置；刻画数据的抽取和转换规则；存储和数据仓库主题相关的各种商业信息。元数据是整个数据仓库运行的基础，如数据的修改、跟踪、抽取、装入和综合等都是依赖于元数据的。故而，有效管理数据仓库的一个重要前提就是拥有描述能力强、内容完善的元数据。元数据可分为四类：关于数据源的元数据、关于数据模型的元数据、关于数据仓库映射的元数据以及关于数据仓库使用的元数据。

表2.1所示为一个元数据的例子，它定义的是数据仓库中的一个表。

表2.1　元数据例表

Table	逻辑名	会员
	定义	在网上书店购买的书籍的个人或组织
	物理存储	DB.table（数据库表）
	表编辑程序名	VALCSTMR（程序名）

6. 数据模型

不同于数据库的是，数据的多维视图是数据仓库中存储的数据模型，它对前端工具、数据仓库的设计和OLAP的查询引擎有直接的影响。

在多维数据模型中，有一部分由一组提供测量值上下文关系的“维”来决定的数据测量值（如销售量、产量、利润等）。以销售量为例，它与销售时间、销售区域和产品名称

相关，由这些相关的维惟一地决定了这个销售测量值（如2013年奥迪在中国的销售量为50万）。因此，将数据测量值存放在由层次的维构成的多维空间所构成的图就是多维数据视图。在多维数据模型中，还可以对一个或多个维做集合运算，例如，按省份和季度对销量进行计算和排序，可以看出不同省份、不同季度的销售情况。一般情况下，时间维对决策中的许多分析都很重要，它是一个具有特殊意义的维度。

可以使用不同的存储机制以及表示模式来实现逻辑上的多维数据模型。一般使用的是星型模型、雪花模型、星网模型和第三范式等多维数据数据模型。

（1）星型模型

星型模型是大部分数据仓库经常采用的一种模型。事实表是星型模式的核心，其他的维表则围绕这个核心表呈现星型分布。维表只与事实表关联，与其他维表则没有任何联系，每个维表只能有一个主码，且该主码同时作为与事实表连接的外码被放在事实表中。事实表中存放了大量的事实数据，且非规范化程度非常高，如在相同的表中出现多个时期的数据。描述性数据存放在维表中。图2.3所示是一个星型模式实例。

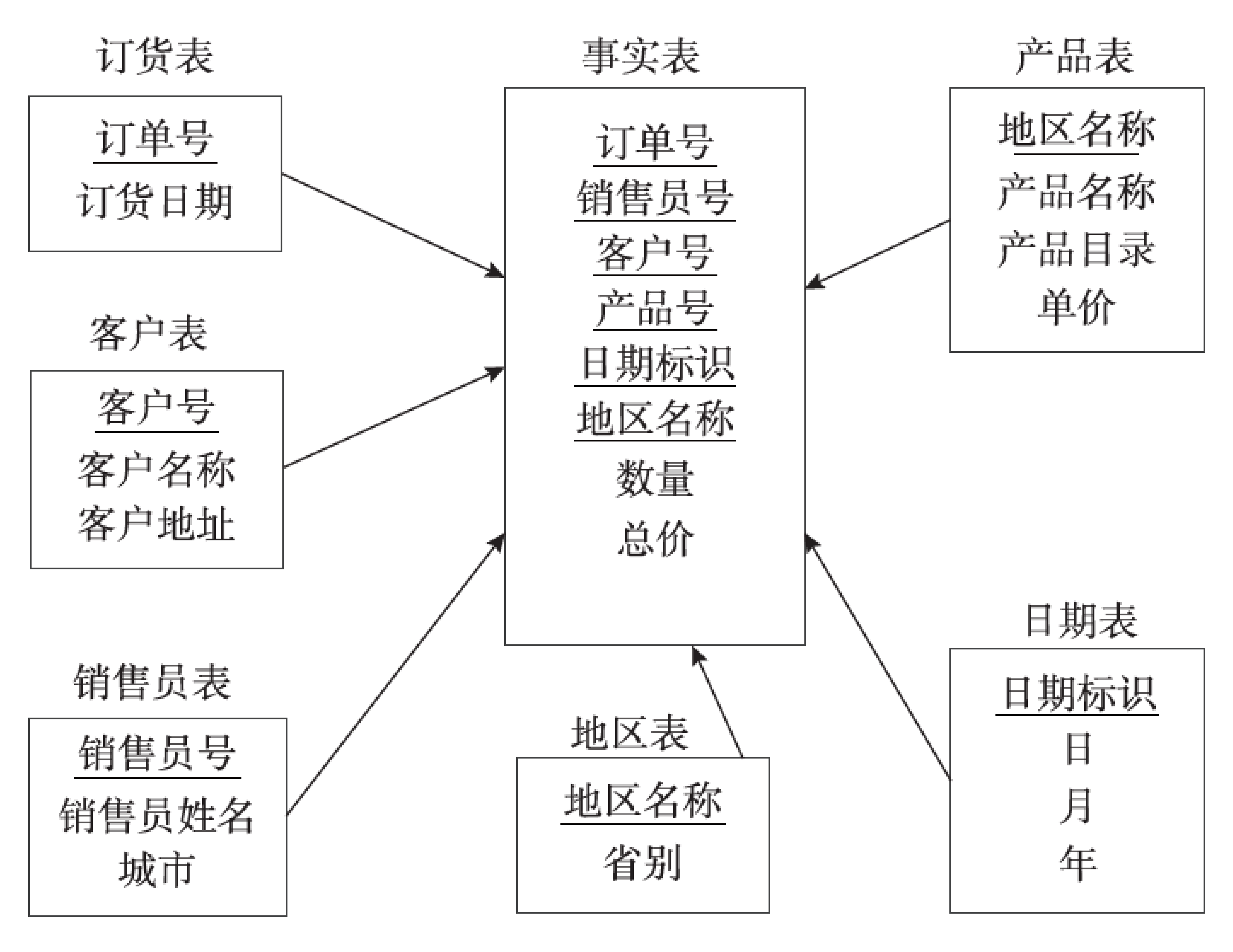


图2.3　星型模型实例

事实表中有大量的记录，而维表中则只有较少的记录。由于针对各个维作了大量的预处理（按照维进行预先统计、分类和排序等），星型模型的数据存取速度较快。例如，预先根据汽车的型号、销售地区以及时间进行销售量的统计，在制作报表时，速度就会加快。

与完全规范化的关系设计相比较，星型模型有着一些明显的差异，它使用大量的非规范化数据，以潜在的存储空间为代价来优化速度。因此星型模型存在很大的数据冗余，因此不适合用于数据量大的情况。此外，星型模型限制了事实表中数量属性的个数，当业务问题发生变化，原来的维不能满足要求时，就需要增加新的维，这种维的变化带来的数据变化是非常复杂且耗时的，因为事实表的主键是由所有维表的主键构成的。

（2）雪花模型

通过将星型模型的维表更进一步地层次化便得到了雪花模型，将原来的各维表扩展成小

的事实表，从而产生一些局部的“层次”区域。它的优点是使得数据的存储量极大限度地降低了，同时为改善了查询性能，它还将较小的维表联合在了一起。

雪花模型的这种方式使系统更进一步专业化和实用化，同时也加大了一些查询的复杂性和用户必须处理的表的数量，而使系统的通用程度下降。数据仓库利用前端工具把用户的需求转变成雪花模型的物理模式，从而完成对数据的查询。

在上面的星型模型中，分别对“产品表”“日期表”“地区表”进行扩展，形成雪花模型数据，如图2.4所示。

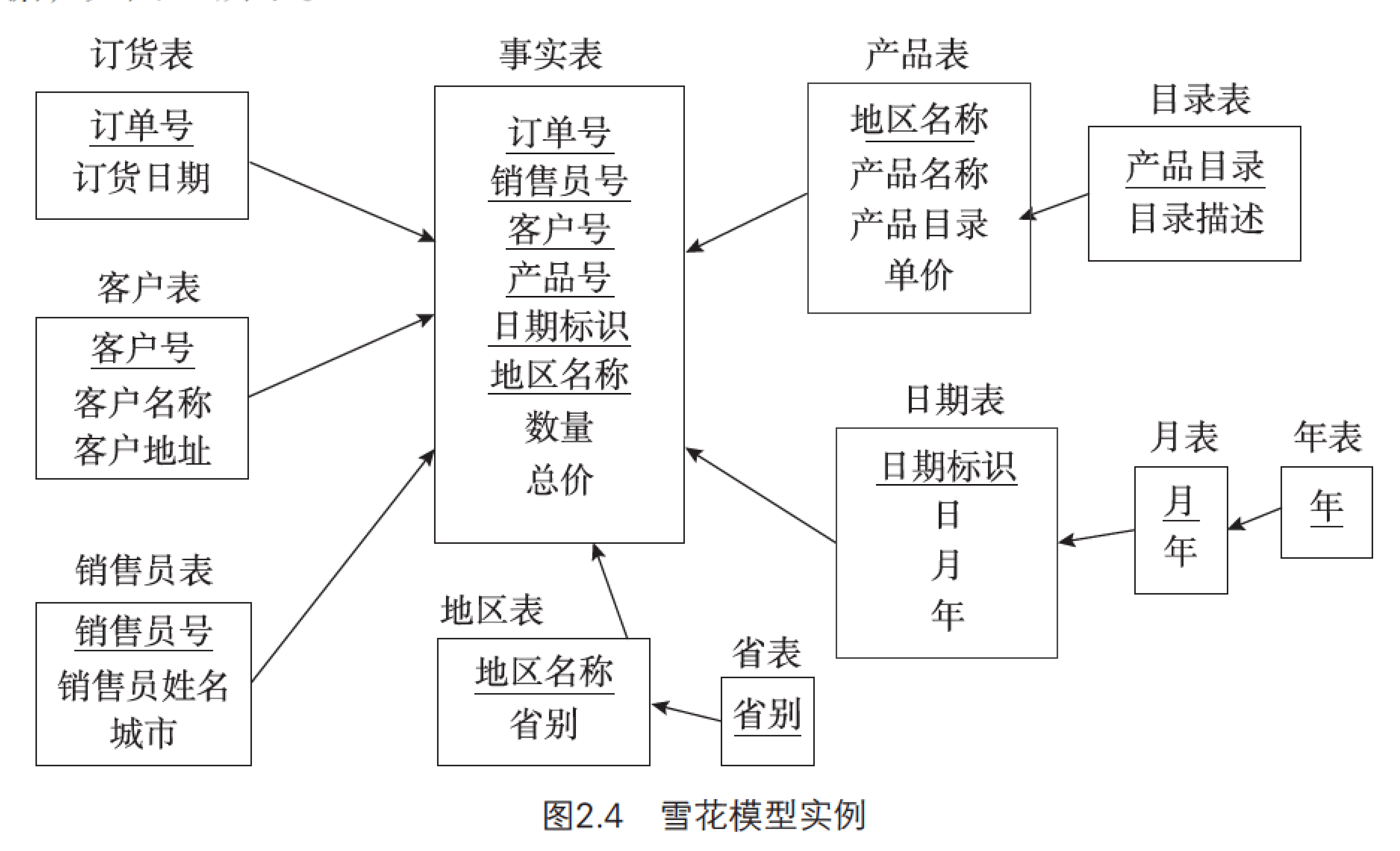


图2.4　雪花模型实例

（3）星网模型

将多个星型模型连接起来形成的网状结构就是星网模型。通过相同的维（如时间维），多个星网模型可以连接多个事实表。

（4）第三范式

范式是符合某一种级别的关系模式的集合，是关系型数据库的构造过程中必须遵循的规则。目前关系型数据库主要有6种范式：第一范式（1NF）、第二范式（2NF）、第三范式（3NF）、第四范式（4NF）、第五范式（5NF）以及第六范式（6NF）。各种范式之间的联系是：第二范式是在第一范式的基础上建立起来的，比第一范式满足更多的条件，第三范式则是在第二范式的基础上满足更多的条件建立的，以此类推便得到各类范式之间的关系。一般情况下，数据库要求满足第三范式就可以了。

第三范式（3NF）指的是这样一种关系，即关系模式中的所有非主属性对任何候选关键字都不存在传递依赖。也就是说，第三范式要求一个数据库表不包含已在其他表中包含的非关键字信息。例如：关系Student(Sno,Sname,Dno,Dname,Location)，关系中的各个属性分别表示学号、姓名、系院号、系院名以及系院地址。其中Sno为关键字，其余各个非主属性完全依赖于Sno这一关键字，所以此关系属于第二范式。关系Student中，对属性Dno，Dname以及Location进行存储、插入、删除以及修改操作时会出现重复的情况，所以此关系存在大量的冗余。造成冗余的原因是此关系中存在传递依赖，即Sno→Dno，Dno→Location，而由于没有

Dno→Sno，所以Sno对Location的决定是通过传递依赖实现的，换句话说，Sno不直接决定非主属性Location。要使每个关系模式中不存在传递依赖，则可以将关系Student分为；两个关系：S(Sno,Sname,Dno)和D(Dno,Dname,Location)，这两个关系属于第三范式。

2.4.2 数据仓库体系结构

数据仓库的不同部分组合在一起就组成了数据仓库的体系结构。体系结构提供了设计开发和部署数据仓库的整体框架结构。本节主要介绍了数据仓库的数据组织结构、系统结构以及运行结构等内容。

1. 数据仓库的数据组织结构

图2.5所示是一个经典的数据仓库的数据组织结构图。

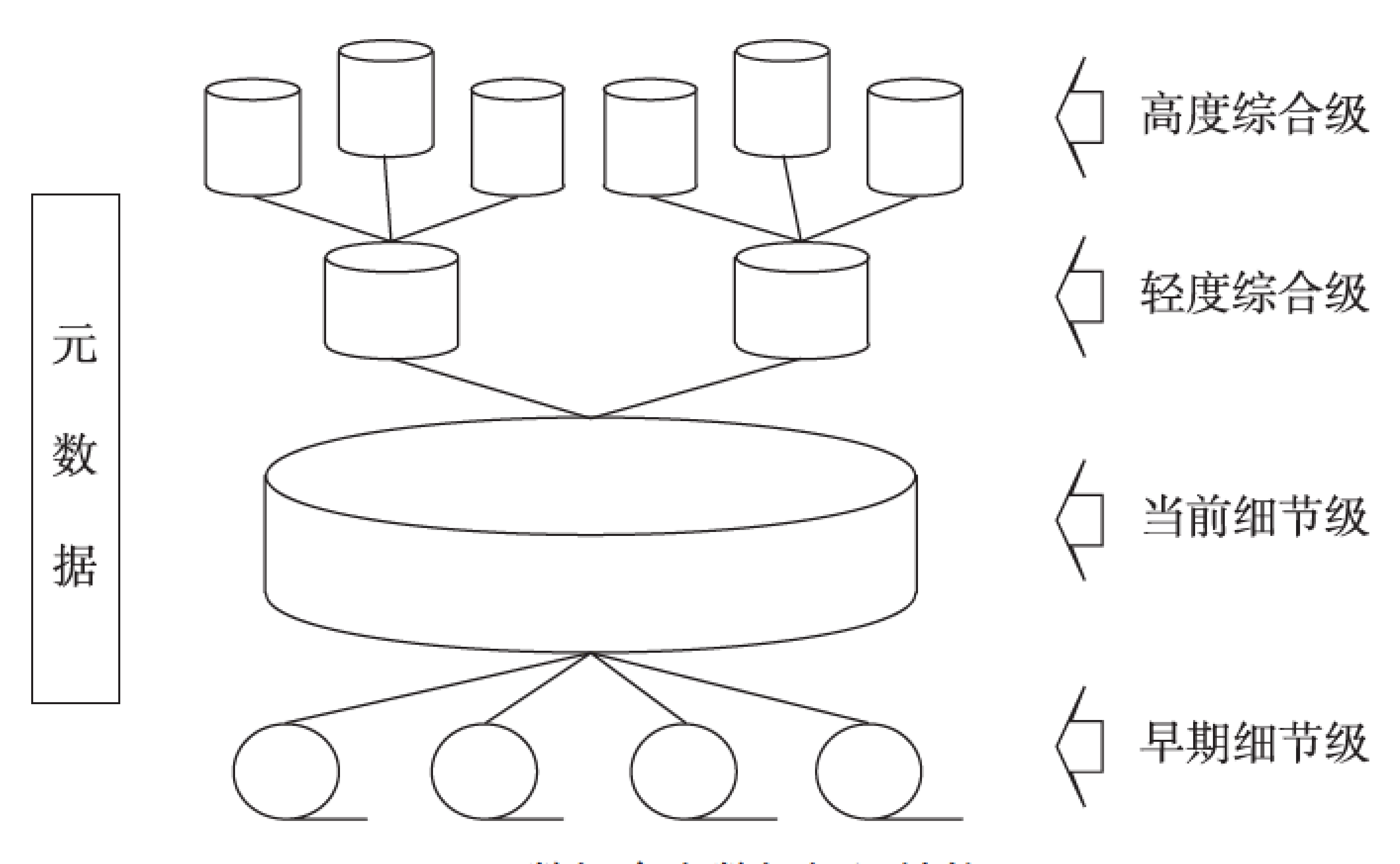


图2.5 数据仓库数据组织结构图

数据在数据仓库中分成早期细节级、当前细节级、轻度综合级以及高度综合级四个级别。一般送入早期细节级的数据都是老化的数据，被综合之后的源数据，先要进入到当前细节级，接着根据具体需要选择进一步地综合从而进入轻度综合级甚至高度综合级。

2. 数据仓库系统结构

数据仓库系统分为三部分，分别是数据仓库管理、数据仓库以及分析工具。数据仓库结构如图2.6所示。

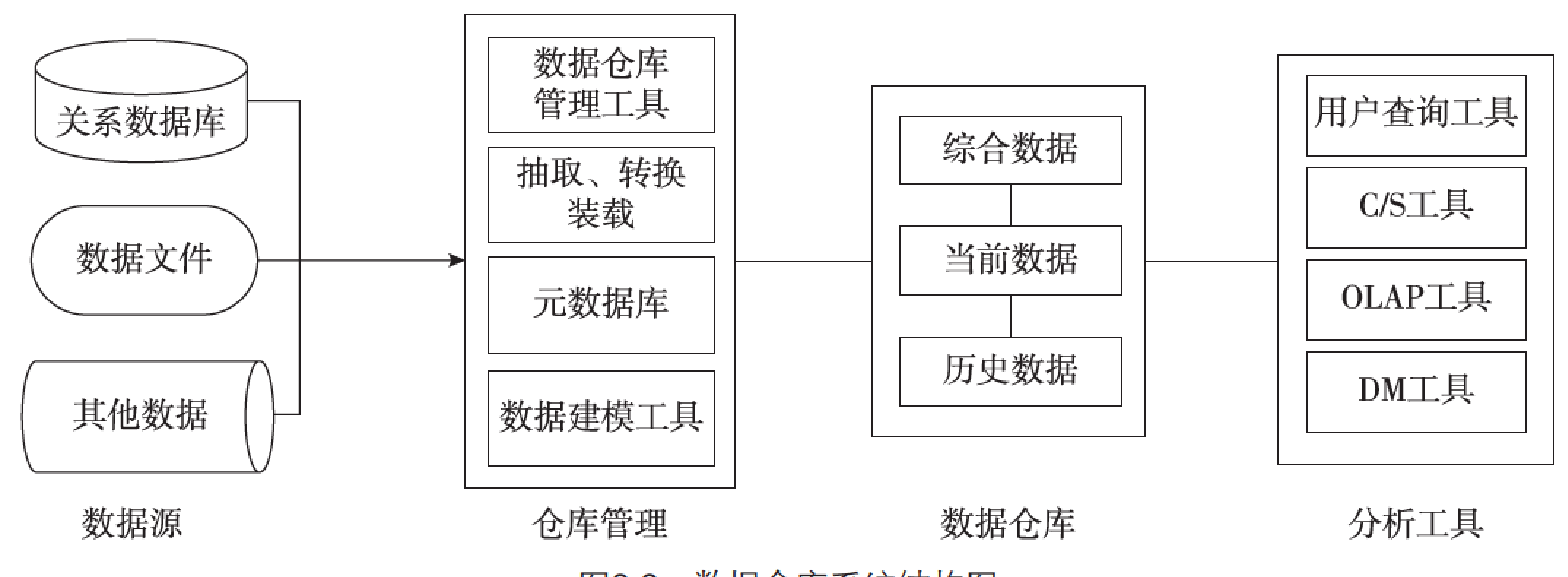


图2.6 数据仓库系统结构图

数据仓库从各种数据源中获取数据。数据源可以是企业内部原有的关系数据库，也可以是企业外部的第三方市场调查报告或者各种文档提供的数据。当明确数据仓库信息是什么需求之后，先要对数据进行建模，接着是确定从源数据到数据仓库的ETL过程，最后进行维数划分和数据仓库的物理存储结构的确定。依靠数据仓库管理系统（DWMS）来完成仓库的管理工作，包括对数据的安全、备份、归档、维护和恢复等。数据仓库管理系统由定义部件、数据获取部件、管理部件、目录部件（元数据）、DBMS部件组成。

由于数据仓库的数据量很大，为了能够从数据仓库中获得能够辅助决策的信息，完成决策支持系统的多项要求，一套具有强大功能的分析工具对于数据仓库来说必不可少。目前分析工具集主要包括两类：查询工具以及挖掘工具。数据仓库的查询通常是指对分析要求的查询，一般包含可视化工具和多维分析（OLAP）工具两种。

3. 数据仓库运行结构

数据仓库应用的运行结构是经典的客户/服务器（C/S）形式。数据仓库采用的是服务器结构的形式，服务器端需要完成多种辅助决策的SQL查询、复杂的计算以及各类综合功能等。客户端完成的工作主要包括：格式化查询、客户交互、结果显示以及报表生成等。现在，在客户与数据仓库服务器之间增加一个多维数据分析（OLAP）服务器的三层C/S结构形式越来越普遍，如图2.7所示。

图2.7　数据仓库应用的三层C/S结构

这种三层结构形式使数据仓库应用工作效率更高，位于客户端和数据仓库服务器之间的OLAP服务器对原客户端和数据仓库服务器的部分工作进行了集中以及简化，使得决策支持的服务工作被加强与规范化，减少了系统数据的传输量。

2.4.3　数据仓库设计与实施

面向主题的、集成的、不可更新的以及随时间的变化不断变化是数据仓库的特点，这些特点导致传统的数据库开发所使用的设计方法并不适用于数据仓库的系统设计。不同于传统数据库开发的是：数据仓库系统的开发者在最初并不能够确切了解到用户的明确而详细的需求，因为数据仓库系统的原始需求是不明确的，也会不断变化与增加，用户也只能提供部分需求或者是需求的大方向，对将来的需求更是无法确切地预料。为此，运用原型法对数据仓库进行开发是一种合适的方法。原型法是一种先从构建简单的基本框架入手，然后再不断丰富完善整个系统的一种软件开发方法。然而，数据仓库的设计并不等同于寻常意义上的原型法，而通常是由数据来驱动数据仓库的设计。数据仓库的开发主要是在原有的数据库系统基础上开展的，它的目的是对已有的数据库中的数据资源进行有效地抽取、综合、集成和挖掘。原型法与系统生命周期法的主要区别在于数据仓库的开发是一个不断循环、反馈，从而使系统不断增长和完善的过程。图2.8形象地说明了构建数据仓库的这个过程。

虽然如此，数据仓库的设计并非毫无步骤可言，其步骤如图2.9所示。

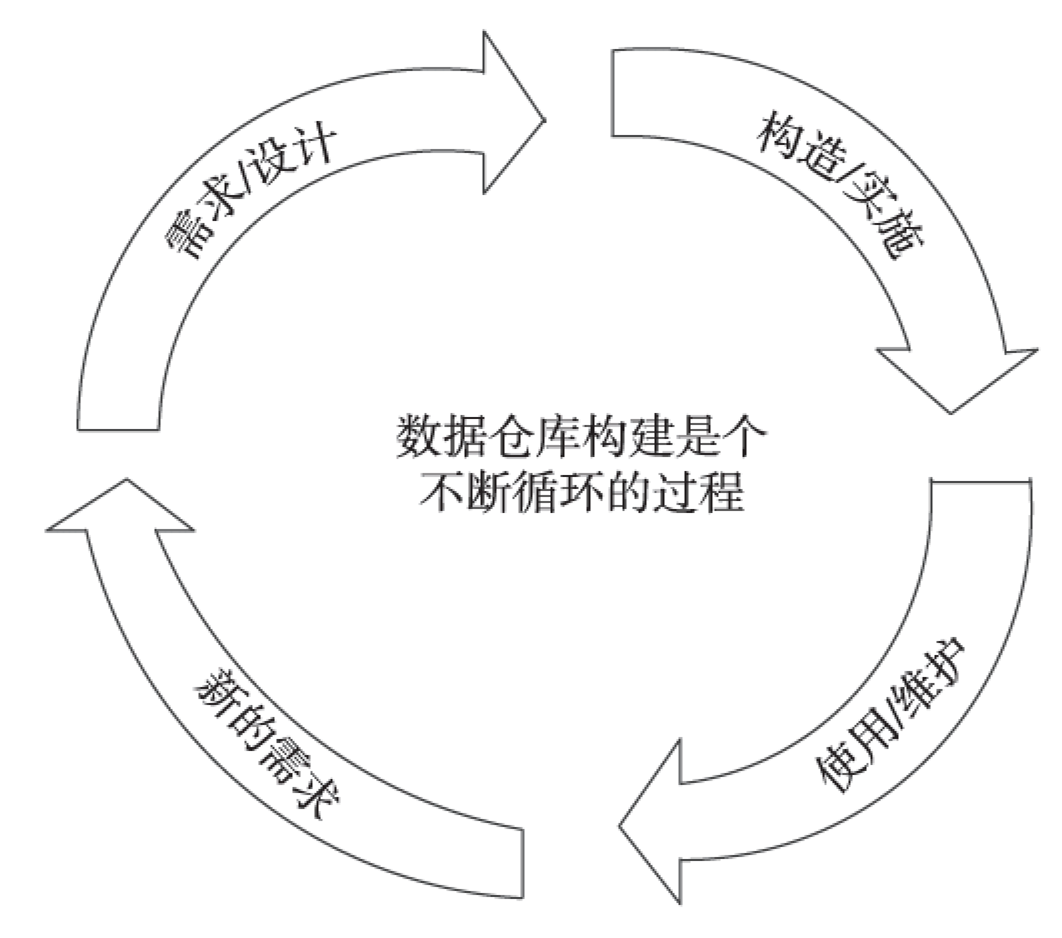


图2.8 数据仓库构建过程图

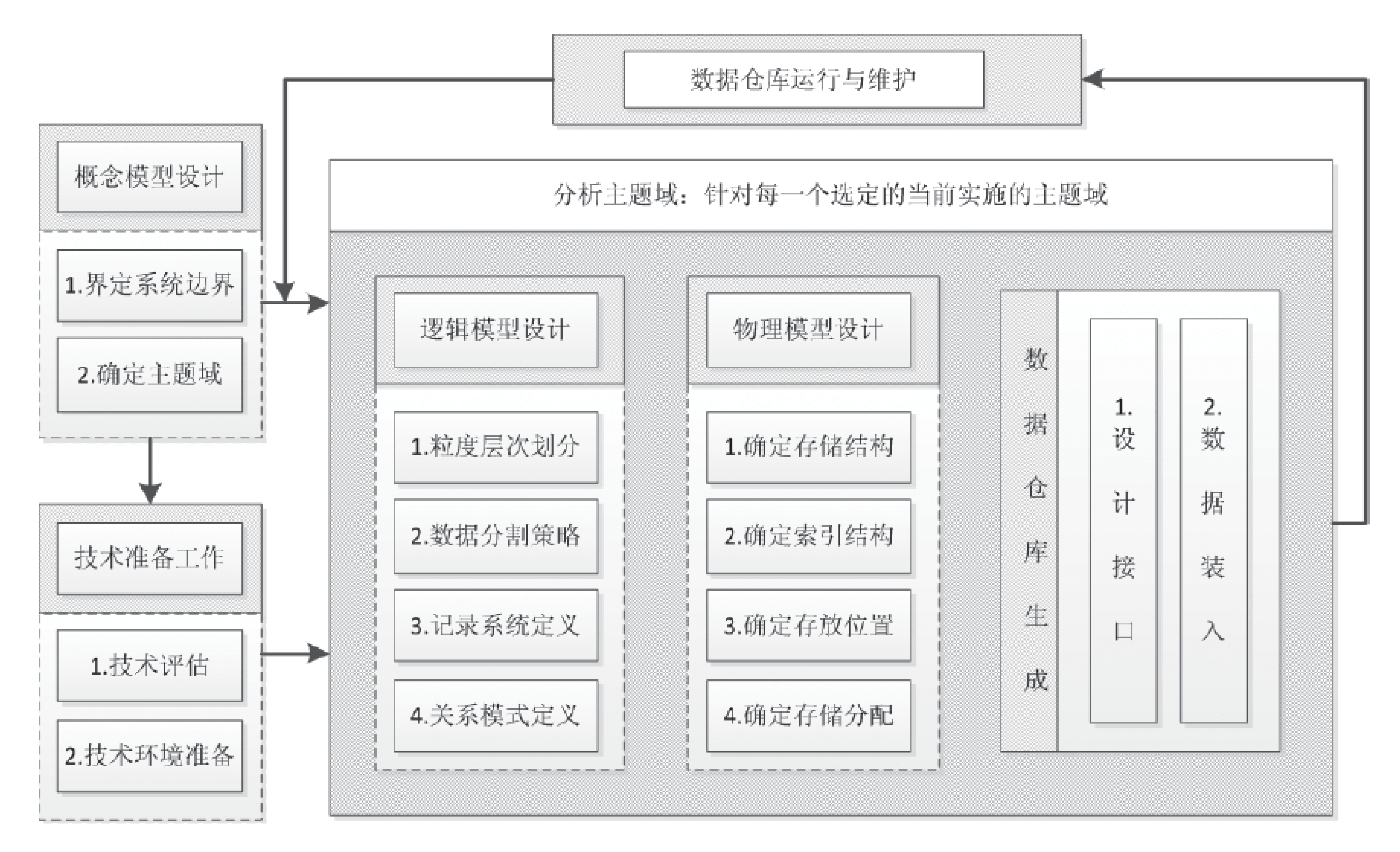


图2.9 数据仓库设计的步骤

需要说明的是，数据仓库设计的步骤并不像上图描述的那么绝对，决策人员与开发者在数据仓库的开发过程中要至始至终共同参与和密切协作，始终以为企业开发数据仓库为目的，保持灵活的头脑，合理安排工作。

下面分别对图2.9所示的六个主要设计步骤进行介绍。

1. 概念模型设计

概念模型是概念型工具，它是连接主观与客观之间的桥梁，是为一定的目标设计系统以及收集信息而服务的。在原有数据库的基础上建立一个较为稳固的概念模型是概念模型设计阶段希望获得的结果。通过集成和重组原有数据库系统中的数据，最终形成的数据集合也就是数据仓库。因此在数据仓库的概念模型设计过程中，如何进行数据仓库系统的概念模型设计并不是首要任务，而是应该先分析和理解原有数据库系统，了解在原数据库系统中有什么、数据如何组织及分布等。首先，要对现有数据库中的内容进行一个全面而清晰的了解，

这一过程是通过查看原来数据库的设计文档和数据字典中的数据库关系模式来实现的；其次，数据仓库是面向全局建立的概念模型，它提供了一个统一的概念视图以集成来自各个面向应用的数据库的数据。概念模型的设计主要是在概念层次上进行的，所以在进行概念模型建立时，具体的技术条件限制并不在考虑范围内。

概念模型设计所包含的工作主要是确定系统边界和系统所包含的主题域。

2. 技术准备工作

管理数据仓库的技术要求与操作型环境中的要求有很大区别，而且两者考虑的方面也不同。通常情况下，都是从操作型数据中将分析型数据分离开，把它们放于数据仓库之中。技术评估和技术环境准备是技术准备阶段的主要工作。这一阶段的主要成果应有：软硬件配置方案、技术评估报告以及系统（软、硬件）总体设计方案。

3. 逻辑模型设计

逻辑模型设计阶段主要进行的工作有：分析主题域，同时确定当前需要装载的主题；确定粒度层次划分以及数据分割策略；定义关系模式和记录系统。通过定义每一个当前要装载的主题的逻辑来实现逻辑模型设计的成果，且在元数据中记录相关内容，其中包括：适当的粒度划分以及表划分、定义合适的数据来源、合理的数据分割策略等①。

4. 物理模型设计

物理模型设计的主要工作是对数据的存储结构和存放位置进行确定，并明确索引策略及存储分配。要确保数据仓库物理模型的实现，对设计人员应有以下几方面的要求。

（1）能够对所选用的数据库管理系统有全面的了解，确定数据的存储结构与存取方法。

（2）能够对影响系统时间和空间效率的平衡和优化的重要依据有所了解，诸如数据使用频度、使用方法、数据规模、数据环境和响应时间要求等。

（3）确定索引策略，索引一旦建立就几乎不需要维护，但是在建立专用的、复杂的索引时却要付出一定的代价。

（4）确定存储分配及数据存放位置，对外部存储设备的特性（如分块原则、块大小的规定、设备的I/O特性等）有所了解。

5. 数据仓库生成

在两个环境不相同的记录系统之间建立一个接口，通过这个接口，才能将原数据库中的数据装载到数据仓库环境中。该接口应具备的功能包括：有效扫描现有记录系统，从面向应用和操作的环境中生成完整的数据，基于时间将数据进行转换、清洗、集成及更新数据。在接口设计完成之后，要通过运行接口程序，将数据装入到数据仓库中去。

6. 数据仓库使用和维护

在数据仓库建立完成，数据被加载到数据仓库之后，还需要对数据仓库进行后续的使用

---

① http://wenku.baidu.com/link?url=grfH--xV5jB-Q_1o2Dw-mB0VbP8sHHaPBdL0t_TPJWAXJEC2eHqY91rTMz2oPCQ7d5bsuiO-WONS9yetRNP29rctVM-4OoAr-dO8ULuoEqq

进行管理和维护。一方面，要在数据仓库中建立起DSS应用以使数据仓库中的数据能够服务于决策分析；一方面，开发人员可以根据用户的使用情况以及反馈得到的新需求，对系统作进一步的完善；另一方面，要对数据仓库中的一些日常活动进行管理，例如对粒度级别进行调整、管理元数据、对数据仓库当前的具体数据进行刷新、将过时的数据转变为历史数据、不再使用的数据清除掉等。

2.4.4 数据抽取、转换和装载

数据仓库中的数据来自多种业务数据源。不可避免地，不同原始数据库中的数据来源、格式是不一样的，因而在系统实施、数据整合过程中会出现一系列问题，必须经过抽取、转换和装载的过程，才能把数据库中的数据真正存储到数据仓库中去，这个过程就是ETL过程。ETL过程将对源系统中的相关数据进行改造，使它们变成有用的信息存储在数据仓库中。不能对源数据进行正确地抽取、清洗和用正确的格式进行整合，就没有数据仓库中的战略信息，也就不能进行数据仓库的查询处理功能。

1. ETL概述

ETL是用来实现异构多数据源的数据集成的工具，是数据仓库、数据挖掘和商业智能等技术的基石。

ETL工具的功能包括：

- 数据的抽取。将数据从不同的网络、不同的操作平台、不同的数据库及数据格式、不同的应用中抽取出来。
- 数据的转换。数据转换（数据的合并、汇总、过滤、转换等）、重新格式化和计算数据、重新构建关键数据以及总结与定位数据。
- 数据的装载。将数据跨网络、操作平台装载到目标数据库中。

ETL的每一个部分都要达到一个重要的目标，每个功能都非常重要。由于源系统的性质，这对ETL提供的功能也很具有挑战性。源系统种类繁多，彼此差异较大，通常ETL需要应对多个平台上的操作系统，而且很多源系统都是采用过时技术的陈旧应用系统。对数据仓库而言，至关重要的是历史数据，这些数据往往是不被保存在操作型系统中的。很多旧系统中的数据质量也各不相同，需要花费大量的时间进行处理。源系统之间的数据普遍缺乏一致性，随着时间的变化，数据结构也可能会发生变化。

在整个项目中，ETL功能设计、测试和部署不同处理过程会占用很大一部分工作量。源系统的性质和复杂程度使得数据抽取本身很复杂，源系统的原数据包含源系统中每一个数据库和每一个数据结构的信息。在数据转换过程中，要应用多种形式的转换技术，必须重新定义内部数据结构，对数据重新排序，应用多种形式的转换技术，给缺失值增加新的默认值，设计性能优化所需要的所有聚集。最初的装载工作可能会往数据仓库中存入数以百万计的数据行，有时可能会花两周甚至更多的时间来完成最初的物理装载。总之，数据的抽取、转换、装载都是费劲且耗时的工作。

ETL过程的主要步骤如图2.10所示。

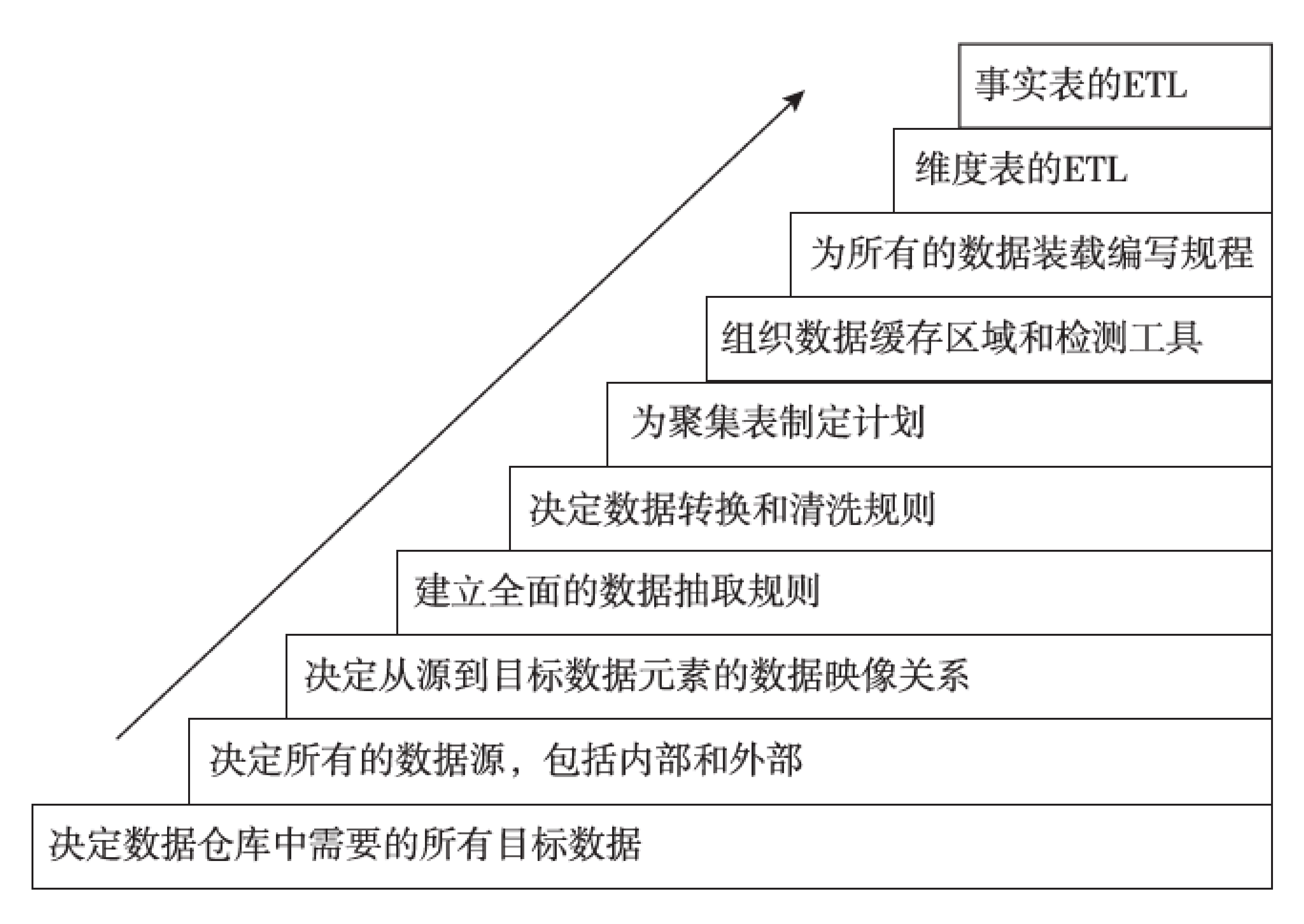


图2.10　ETL处理过程的主要步骤

2. 数据抽取

数据抽取就是一个从数据源中抽取数据的过程。具体来说，就是搜索整个数据源，使用某些标准选择合乎要求的数据，并把这些数据传送到目标文件中。对于数据仓库来说，必须根据增量装载工作和初始完成装载的变化来抽取数据。对于操作型系统来说，则需要一次性抽取和数据转换。这两个因素增加了数据抽取工作的复杂性，而且，也促使在内部编写代码和脚本的基础上，使用第三方数据抽取工具。使用第三方工具往往会比内部编程更贵，但是它们记录了自己的元数据，另一方面，内部编程增加了维护的成本，当源系统变化时，也很难维护。而第三方的工具则提供内在的灵活性，只需要改变它的输入参数就可以了。

数据仓库的成功首先取决于有效的数据抽取，所以需要对数据仓库的数据抽取策略的制定给予特别关注。数据抽取的要点包括：确认数据的源系统及结构；针对每个数据源定义抽取过程（人工抽取还是基于工具抽取），确定数据抽取的频率，表示抽取过程进程的时间窗口；决定抽取任务的顺序；决定如何处理无法抽取的输入记录。

通常，源系统的数据是以两种方式来存放的：当前值和周期性的状态。源系统中的大多数数据都是当前值类型，这里存储的属性值代表的是当前时刻的属性值，但这个值是暂时的，当事务发生时，这个值就会发生变化。周期性的状态指的是属性值存储的是每次发生变化时的状态。对于这个类型的操作型数据，进行数据抽取工作会相对容易很多，因为其变化的历史存储在源系统本身当中。

从源操作系统中抽取的数据主要有两种类型：静态数据和修正数据。静态数据是在一个给定时刻捕获的数据，就像是相关源数据在某个特定时刻的快照。对于当前数据或者暂时的数据来说，这个捕获过程包括所有需要的暂时数据。对于周期性数据来说，这一数据捕获包括每一个源操作型系统中可以获得的每个时间点的每一个状态或者事件。在数据仓库的初始装载时一般使用静态数据捕获。修正数据也称为追加的数据捕获，是最后一次捕获数据后的修正。修正数据可以是立刻进行的，也可以是延缓的。在立即型的数据捕获中，有三种数据抽取的方法：通过交易日志捕获、从数据库触发器中捕获或者从源应用程序中捕获。延缓的数据抽取有两种方法：基于日期和时间标记的捕获和通过文件的比较来捕获。

3. 数据转换

抽取后的数据是没有经过加工的，这些数据的质量并没有像数据仓库要求的那样好，是不能直接应用于数据仓库的，必须将所有抽取的数据转换为数据仓库可以使用的数据。数据转换的一个重要任务就是提高数据质量，包括补充已抽取数据中的缺失值，去除脏数据，修正错误格式等。

数据转换功能包含一些基本的任务：选择、分离/合并、转化、汇总和丰富。转换功能主要完成格式修正、字段的解码、计算值和导出值、单个字段的分离、信息的合并、特征集合转化、度量单位的转化、日期/时间转化、汇总、键的重新构造等工作。

由于数据转换的复杂性和涉及范围广，仅靠手工操作是难以完成的，因此，使用转换工具是一种有效的方法。使用转换工具的主要优点就是在数据转换过程中，转换参数和规则都会作为元数据被工具存储起来，这些元数据就会成为数据仓库整个元数据集合的一部分，可以被其他部分共享。尽管转换工具的理想目标是排除手工的方法，但是在实际工作中这却是不可能实现的，即使有最精良的转换工具组合，也要准备好使用内部开发的程序。使用转换工具和手工方法两者结合才是最好的办法。

4. 数据装载

数据装载是指在将数据最终复制到数据仓库之前，把它们复制到一个中间位置。数据仓库的装载工作需要大量的时间，理想状况下，应在操作系统不忙时进行数据的复制，并确保了解自己的商务及其支持系统。确保已经完成了大量的更新，否则不应进行数据的移动。如果数据仓库中的数据来自多个相互关联的操作系统，就应该确保在这些系统同步工作时移动数据。

为了能够高效和及时地把数据装载到数据仓库中，一般都要利用选定的批量装载程序。批量装载程序一般应包括的功能有：按索引对文件进行排序、数据类型转换和操作以及数据统计。

5. ETL工具

ETL工具所要完成的工作主要包括三个方面。首先，在数据仓库和业务系统之间搭建起一座桥梁，确保新的业务数据能够源源不断地进入数据仓库。其次，用户的分析和应用能够反映最新的业务动态，虽然ETL在数据仓库架构的三部分中技术含量并不高，但其涉及到大量的业务逻辑和异构环境，因此在一般的数据仓库项目中，ETL部分往往会消耗最多的精力。最后，从整体角度来看，ETL的主要作用是为各种基于数据仓库的分析和应用提供统一的数据接口，屏蔽复杂的业务逻辑，而这正是构建数据仓库最重要的意义所在。ETL工具的正确选择，可以从多方面考虑，如ETL对平台的支持、对数据源的支持、数据转换功能、管理和调度功能、集成和开放性、对元数据管理等功能出发。

随着各种应用系统数据量的飞速增长，以及对业务可靠性等要求的不断提高，人们对数据抽取工具的要求也在不断提高。比如往往要求对几十，上百个GB的数据进行抽取、转换和装载工作，这种挑战毋庸置疑会要求抽取工具对高性能的硬件和主机提供更多支持。因此，从数据抽取工具支持的平台，可以判断出它能否胜任企业的环境，目前主流的平台包括SUN Solaris、HP-UX、IBM AIX、AS/400、OS/390、Sco UNIX、Linux和Windows等。

由于对数据抽取的要求越来越高以及专业ETL工具的不断涌现，ETL过程早已不再是一个

简单的小程序就可以完成的。目前主流的工具都采用多线程、分布式、负载均衡、集中管理等高性能、高可靠性与易管理和扩展的多层体系架构。

专业的ETL厂商和主流工具主要有：OWB（Oracle Warehouse Builder）、ODI（Oracle Data Integrator）、Informatic PowerCenter（Informatica公司）、AICloudETL、DataStage（Ascential公司）、Repository Explorer、Beeload、Kettle、DataSpider、ETL Automation（NCR Teradata公司）、Data Integrator（Business Objects公司）和DecisionStream（Cognos公司）。

6. ETL展望

ETL有着广阔的发展空间，只有基于数据ETL，数据仓库、数据挖掘以及商业智能等技术才能更好地实现，从而为企业提供决策与预测的基本素材。伴随着现实需求的强劲推动， ETL逐渐成为当前信息技术最活跃的研究领域之一，呈现出通用化、高效化、智能化这三大发展趋势。

企业不管是进行当前事务处理，还是未来预测，其前提就是数据，而提供综合且高品质的数据正是ETL的目的，因此它必然要为众多的高层信息系统提供服务，成为企业各类应用的基础。只有具备良好通用性的ETL软件才能占领未来市场，为此，对未来的ETL软件提出以下几点要求：能够跨网络、跨平台使用；能够支持尽可能多的数据库管理系统（DBMS）、文件系统和数据采集、处理系统；具备良好的可扩展性，对于新的应用能够通过预订的应用程序接口（API）或标准化语言接口编程，以较小的代价实现互连。元数据的标准化、程序逻辑与数据的统一化等相关技术的发展，为ETI提高通用性提供了的动力。

由于针对的是海量数据， ETL对效率极为重视，未来的ETL工具将是高效的数据集成工具。高度的可伸缩性是其必备条件之一，不管是在昂贵的主机系统上，还是在工作站或PC机上，都能够运行ETL。为了能够真正避免重复集成，更加出色、高效地抽取、加载和清洗算法，增量的ETL算法将成为主流。此外，采用并行算法、集群计算、网络运算的ETL工具将领导潮流，为ETL提供廉价高效的计算资源，提高计算性价比。

高度的智能也是未来ETL必备的特征之一。此处将广泛应用专家系统、机器学习、神经网络、人工智能（AI）技术等领域的成果，由机器智能来完成数据源管理、ETL规则定制、数据质量保证等工作。这会在很大程度上减轻了用户的工作量，很多枯燥而繁重的数据集成工作将由ETL来完成。ETL工具的使用也会不断简化，通过运用智能工具，普通用户也能轻松而高效地完成数据的集成与清洗工作。

决定数据仓库能否获取高质量数据的核心是ETL工具，利用ETL工具能够解决各种应用数据零散分布、品质低下的现状，将各种异构信息根据决策需求集中到数据仓库中。待集成多数据源的异构性成为ETL最大的挑战，为了降低系统实现的难度，将数据转化的逻辑规范和物理实现分开管理，通常要把实施ETL过程划分为模式集成与数据集成两个阶段。

2.4.5　联机分析处理

数据仓库系统包括数据仓库层、工具层和它们之间的相互关系。数据仓库系统是由多种技术组成的综合体，主要包括数据仓库、数据仓库管理系统以及数据仓库工具这三个部分。在整个系统中，数据仓库是信息挖掘的基础，处于核心地位；数据仓库管理系统则是整个系统的引擎，承担管理整个系统运转的责任；而整个系统发挥作用的关键是数据仓库工具，数

据仓库唯有采用高效的工具才可以真正将其数据仓库的作用发挥出来。

数据仓库的目标决定了一般的查询工具无法满足数据仓库真正的需求，它需要拥有分析功能更加强大的工具。此处的查询，并不只是对记录级数据的查询（可能会存在此类查询，但绝不会多），更多的是对分析结果（发展趋势或模式总结）的查询，因此对更加友好的表达方式提出要求。例如，为了使用户能更方便、更清晰、更直观地了解复杂的查询结果而采用各种图形和报表工具。数据仓库中最主要的工具是分析型工具。用户或许有各种各样的方式从数据仓库采掘信息，然而大致上都可以分为验证型（verification）和发掘型（discovery）这两种模式。验证型指的是用户利用各种工具通过反复的、递归的检索查询，对自己先前提出的假设进行验证或是否定。多维分析工具是主要的验证型工具。联机分析处理（OLAP）就需要利用多维分析工具。与验证型的工具不同，发掘型的工具并不需要事先提出假设，而是直接从海量数据中发现数据模式，从而预测趋势和行为。发掘型的工具主要是指数据挖掘（Data Mining）。

接下来，将介绍OLAP的相关知识，数据挖掘的相关知识将在第6章进行详细地介绍。

1. OLAP的基本概念

OLAP是联机分析处理（On-Line Analytical Processing）的简称，是由关系数据库之父E.F.Codd于1993年提出的。随着市场竞争的日趋激烈，企业更加注重决策的即时性和准确性，E.F.Codd认为终端用户对数据库查询分析的需求早已不满足于联机事务处理（OLTP），用户分析的需求也不满足于SQL对大数据库的简单查询。由于关系数据库不能进行大量计算，所以查询的结果并不能满足决策者提出的需求，导致用户的决策分析无法得到想要的结果。

OLAP委员会对联机分析处理的定义为：从原始数据中转化出来的、能够真实反映企业多维特性，并能够真正为用户所理解的数据称为信息数据。联机分析处理是能够获得对数据更深入了解的一类软件技术，能使分析人员、管理人员或执行人员对信息数据从多种角度进行快速、交互、一致地存取。目前所指的联机分析处理，主要是指对数据的一系列交互查询的过程，这些查询过程要求对数据进行多层次、多阶段的分析处理，以获得高度归纳的信息。从作用上来说，联机分析处理是一种快速软件技术，能够实现多维信息共享，针对特定问题的联机数据访问和分析。OLAP也可以说是多维数据分析工具的集合，其目标是满足多维环境下，对特定的查询和报表需求或决策支持，其中“维”是它的技术核心。

OLAP技术的主要特点有以下两个：一是在线性（On-line），表现为能快速响应和交互操作用户请求；二是多维分析（Multi-Analysis），即是OLAP技术的核心所在。

多维分析是指采用切片、切块、旋转等各种分析动作，对以多维形式组织起来的数据进行剖析，使最终用户对数据库中的数据进行多角度、多侧面的观察，从而更深入地了解包含在数据中的信息和内涵。多维分析方式能够减少混淆及降低错误解释的出现，这是由它迎合了人的思维模式决定的。多维分析的基本动作主要有：切片、切块、上卷、下钻及旋转。

（1）切片（Slice）

在多维数组中选定一个二维子集的动作叫做切片，即从多维数组（维1，维2，…，维n，变量）中选定两个维，维i和维j，在这两个维上选取某一区间或任意维成员，而将其余的维都取定一个维成员，得到的就是多维数组在维i和维j上的一个二维子集。这个二维子集就称为多维数组在维i和维j上的一个切片，表示为：（维i，维j，变量）。如图2.11所示，选

定两个维（“贷款”维度和“经济性质”维度），而在“时间”维度上选定一个维成员（如“第1季度”），就得到了“贷款”和“经济性质”两个维上的一个切片。这个切片表示了在第一季度各经济性质和各贷款类别的贷款总额。

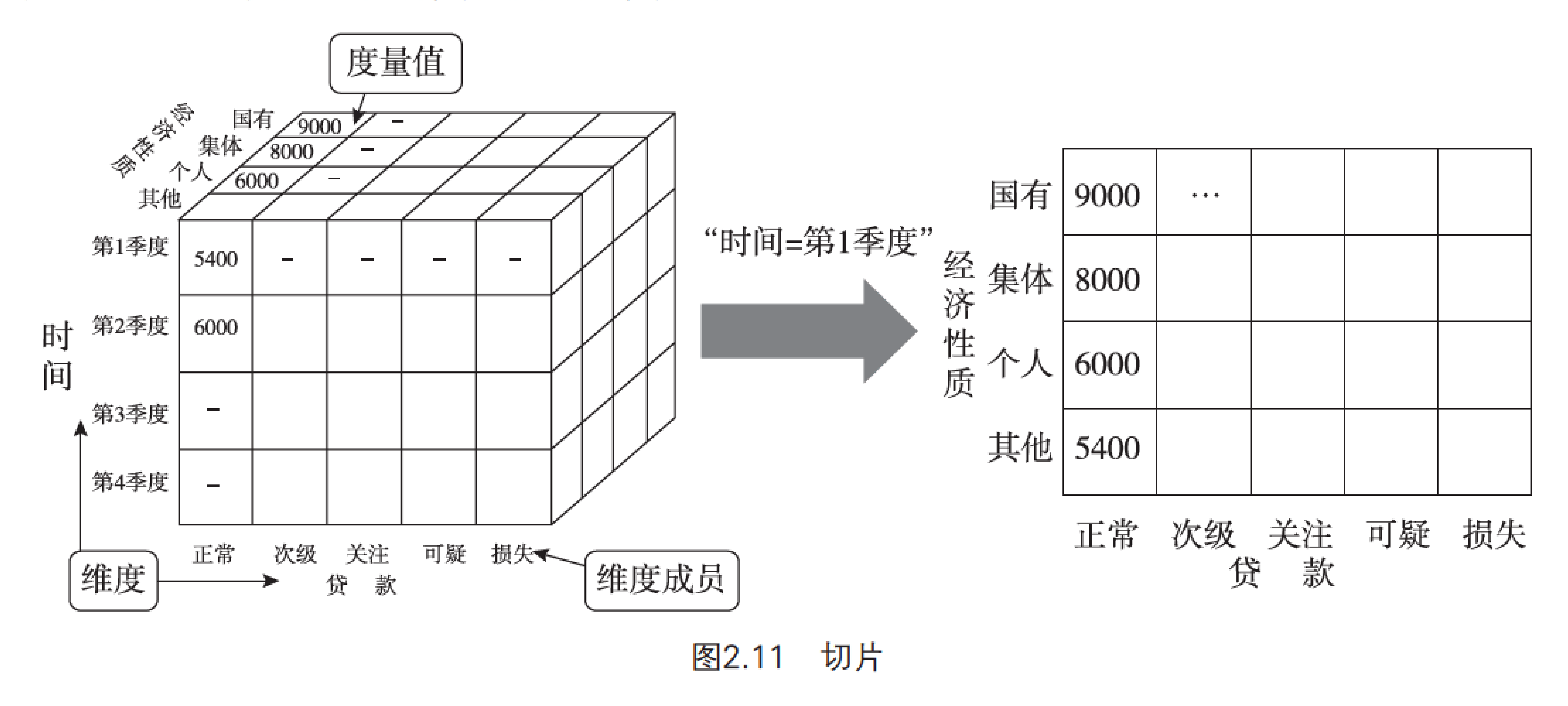


图2.11　切片

（2）切块（Dice）

与切片类似。在多维数组中选定一个三维子集的动作叫做切片，即选定多维数组（维1，维2，…，维n，变量）中的三个维，维i、维j和维r。在这三个维上选取某一区间或任意维成员，而将其余的维都取定一个维成员，则得到多维数组在维i、维j和维r上的一个三维子集，这个三维子集称为多维数组在维i、维j和维r上的一个切块，表示为：（维i，维j，维r，变量）。如图2.12所示，在“时间”维度和“贷款”维度上各选定两个维成员（如“第1季度”和“第2季度”，“正常”和“次级”），在“经济性质”维度选定三个维成员（“集体”“个人”和“其他”）就可以得到一个切块了。

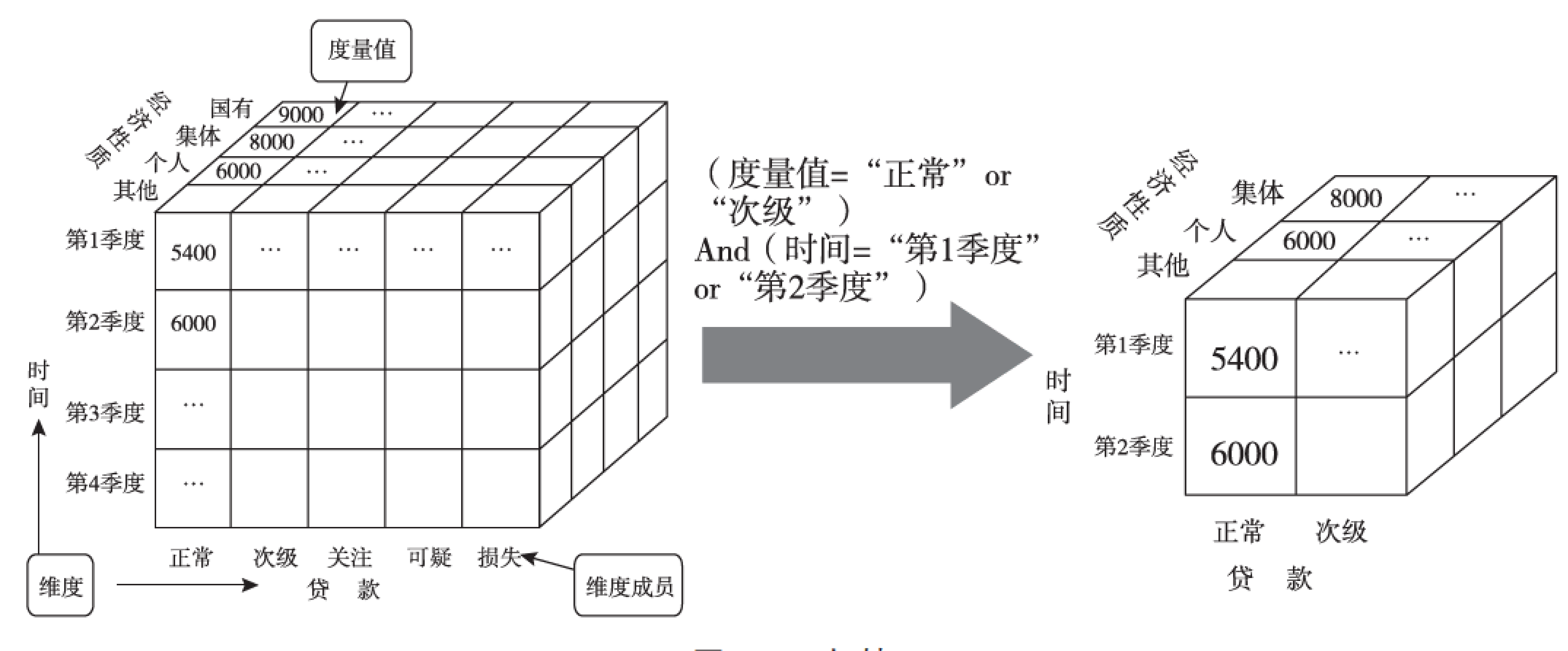


图2.12　切块

（3）上卷（roll up）

在数据立方体中执行聚集操作，通过在维级别中上升或通过消除某个或某些维来观察更概括的数据。沿着时间维上卷，由“季节”上升到半年，如图2.13（a）所示，或者消除“经济性质”这一维度，如图2.13（b）所示，就得到更高层次的汇总数据。

度量值
经济性质 国有 集体 个人 其他
9000 8000 6000
时间 第1季度 第2季度 第3季度 第4季度
5400 6000
维度 维度成员
正常 次级 关注 可疑 损失
贷款
沿着时间维上卷，由“季度”上升到半年
经济性质 国有 集体 个人 其他
14000 12000 10000
时间 上半年 下半年
11400
正常 次级 关注 可疑 损失
贷款

（a）

度量值
经济性质 国有 集体 个人 其他
9000 8000 6000
时间 第1季度 第2季度 第3季度 第4季度
5400 6000
维度 维度成员
正常 次级 关注 可疑 损失
贷款
消除“经济性质”维度
时间 第1季度 第2季度 第3季度 第4季度
28400 21000
正常 次级 关注 可疑 损失
贷款

（b）

图2.13 上卷

（4）下钻（Drill down）

通过在维级别中下降或通过引入某个或某些维来更细致地观察数据，与上卷正好相反。如图2.14所示，沿着时间维下钻，就得到了每个月的各经济性质各贷款类型的贷款更具体的信息。

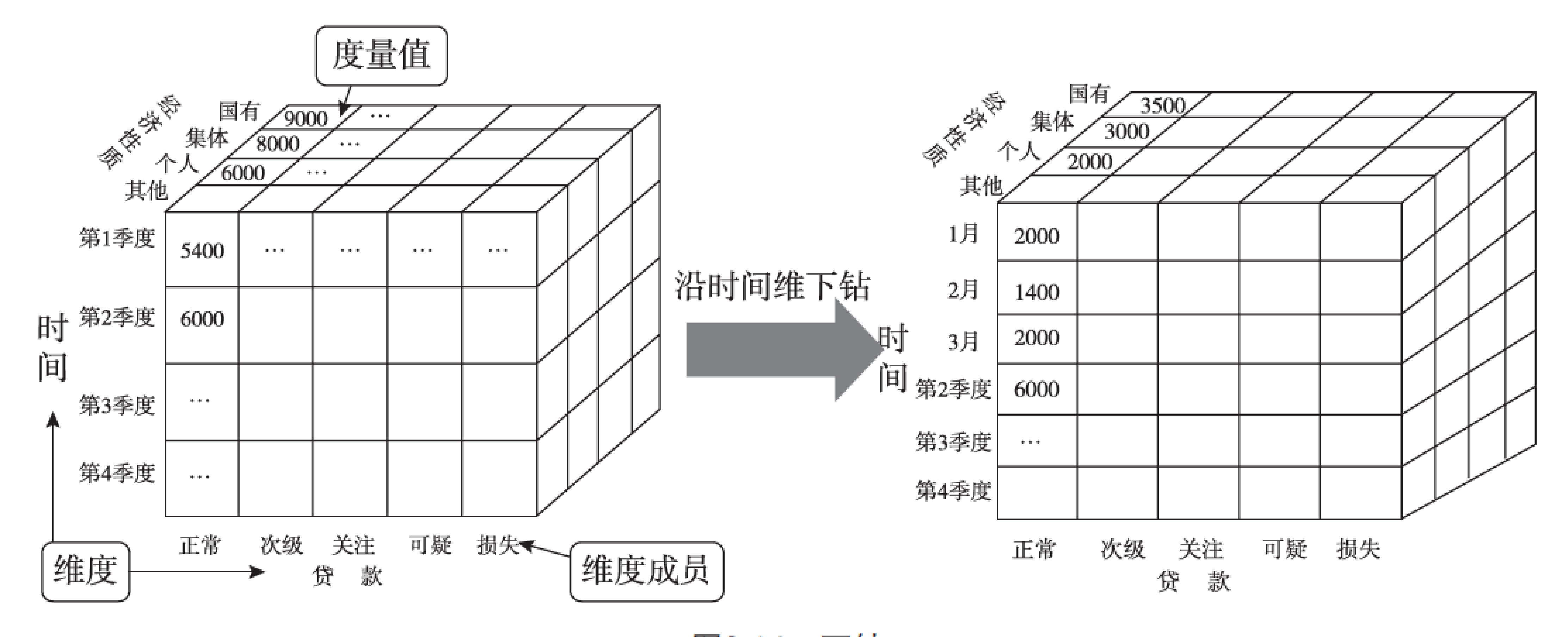


图2.14 下钻

（5）旋转（Rotate）

旋转，即改变一个报告或页面显示的维方向。旋转有以下几种方式：交换行和列、把某一个行维移到列维中去、把页面显示中的一个维和页面外的维进行交换（令其成为新的行或列）。如图2.15所示，将“时间”维和“经济性质”维进行了变换。

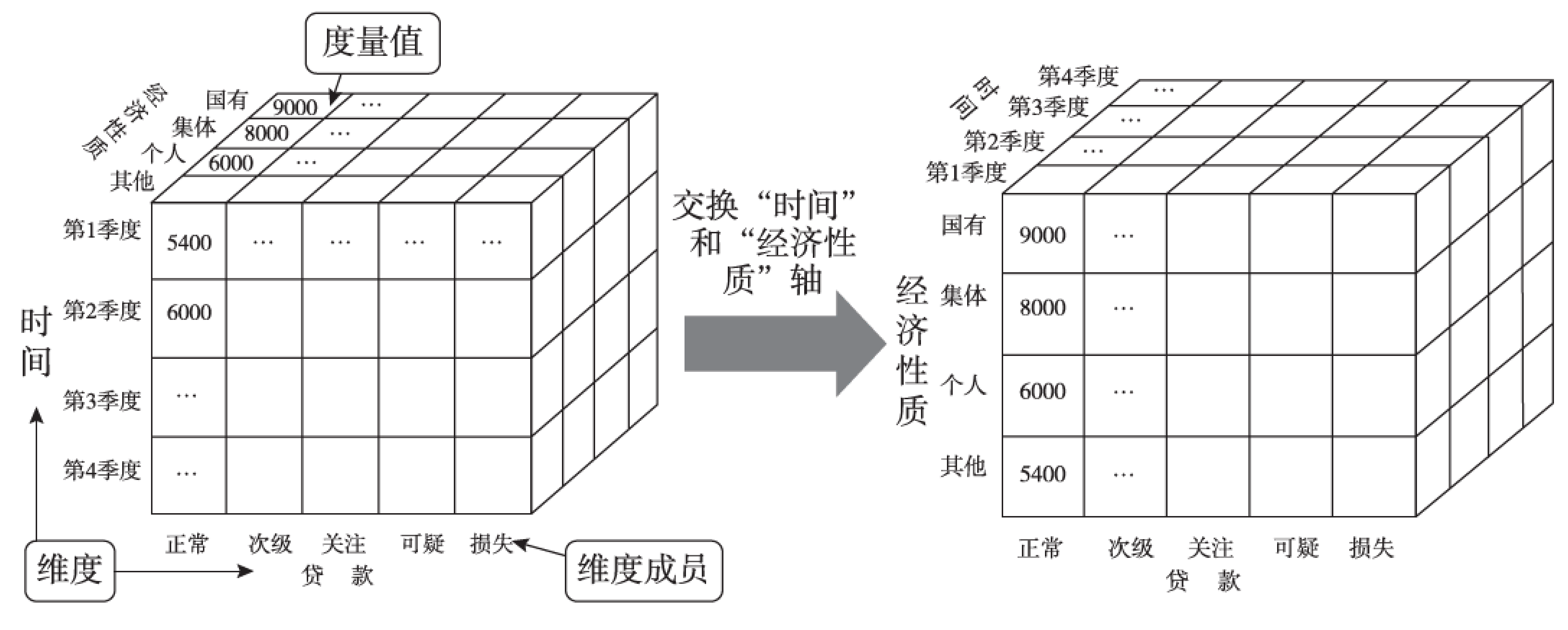


图2.15　旋转

2. OLAP特征及衡量标准

OLAP主要有以下四大特征：

多维概念视图是OLAP最显著的特征。在OLAP数据模型中，将多维信息抽象为一个立方体，其中包括维和度量。维是人们观察数据的特定角度，是考虑问题时的一类属性，而度量表示的是多维数组的取值。多维结构是OLAP的核心，在用户面前OLAP展现的是一幅幅的多维视图。

快速响应用户的分析需求是OLAP的第二大特征。一般认为在几秒内对用户的分析请求做出响应的OLAP系统才是正常的。如果响应时间超过30秒，用户可能就会不耐烦，导致失去分析主线索，从而影响分析质量。因此需要更多诸如大量的事先运算、专门的数据存储格式以及特别的硬件设计等技术上的支持。

OLAP的第三个特征是它的分析功能。与应用有关的任何逻辑分析和统计分析它都应该能处理。用户的数据分析不仅能在OLAP平台上进行，也可以连接到其他工具，如成本分析工具、时间序列分析工具、数据挖掘和意外报警等外部分析工具上。OLAP的基本分析操作有切片、切块、下钻、上卷及旋转。

OLAP的第四个特征是它的信息性。OLAP系统能够及时地获取并管理大容量信息，无论多大的数据量以及数据存储在什么地方。

E.F.Codd给出了十二条基本准则，以便对OLAP产品进行评价。

- 透明性准则；
- OLAP模型必须提供多维概念视图；
- 存取能力准则；
- 客户/服务器体系结构；

- 稳定的报表性能；
- 维的等同性准则；
- 多用户支持能力准则；
- 动态的稀疏矩阵处理准则；
- 非受限制的跨维操作；
- 灵活的报表生成；
- 直观的数据操纵；
- 不受限维与聚集层次。

3. OLAP服务器类型

根据存储器的数据存储格式，OLAP系统可以分为关系型OLAP（Relational OLAP，简称ROLAP）、多维型OLAP（Multi-Dimensional OLAP，简称MOLAP）以及混合型OLAP（Hybrid OLAP，简称HOLAP）三种。

关系数据库是关系型OLAP（ROLAP）的核心，ROLAP将用作分析的多维数据以及根据应用的需要有选择地定义一批实视图作为表存储在其中。只选择那些计算工作量比较大、应用频率比较高的查询作为实视图，而不是把每一个SQL查询都作为实视图保存。为提高查询效率，对具有针对性的OLAP服务器的查询而言，优先选择利用已经计算好的实视图来生成查询结果。这是一种介于关系后端服务器和用户前端工具之间的中间服务器，同时用作ROLAP存储器的RDBMS也针对OLAP作相应的优化。它比MOLAP技术具有更大的可规模性。例如Mircostrategy的DSS和Informix的Metacube都采用了ROLAP方法。

多维型OLAP（MOLAP）是在物理上以多维数组的形式将OLAP分析所用到的多维数据进行存储，然后会产生“立方体”的结构。用多维数组的下标值或下标的范围映射维的属性值，总结数据以多维数组的值的形式在数组的单元中存储。由于MOLAP从物理层起实现，存储结构是新的，因此又称为物理OLAP（Physical OLAP）。相比较而言，ROLAP的物理层仍采用关系数据库的存储结构，主要借助于一些中间软件或软件工具实现，因此也称为虚拟OLAP（Virtual OLAP）。

混合型OLAP（HOLAP）的提出是由于MOLAP和ROLAP的结构完全不同，各自的优点和缺点也不同，分析人员在设计OLAP结构时比较困难。HOLAP则结合了MOLAP和ROLAP两种结构的优点。HOLAP虽然还没有一个正式的定义，但能满足用户各种复杂的分析请求，HOLAP结构不是将MOLAP与ROLAP的结构简单组合，而是将这两种结构技术优点有机地结合了起来。例如微软的SQL Server 7.0 OLAP服务就支持混合OLAP服务器。

4. OLAP的实施

OLAP要对来自基层的操作数据（由数据库或数据仓库提供）进行多维化或预综合处理，因此，它是三层客户/服务器结构的，这与传统OLTP软件的两层客户/服务器结构有所不同。

三层客户/服务器的结构示意图如图2.16所示。它的主要特点是把应用逻辑（或业务逻辑）、DBMS及GUI严格区分开。复杂的应用逻辑主要集中存放于应用服务器上，而不是分布于网络上众多的PC机上，其高效的数据存取由服务器提供，然后安排后台进行处理和报表的

预处理。由三层客户/服务器结构示意图可看出，OLAP实施有两点非常关键：一是OLAP服务器的设计，即如何将来自多个不同数据源或数据仓库的数据进行组织；二是OLAP服务器与前端软件的沟通。多维数据分析就是连接OLAP服务器与前端软件的桥梁，因此，OLAP服务器的构建必须以多维方式进行。

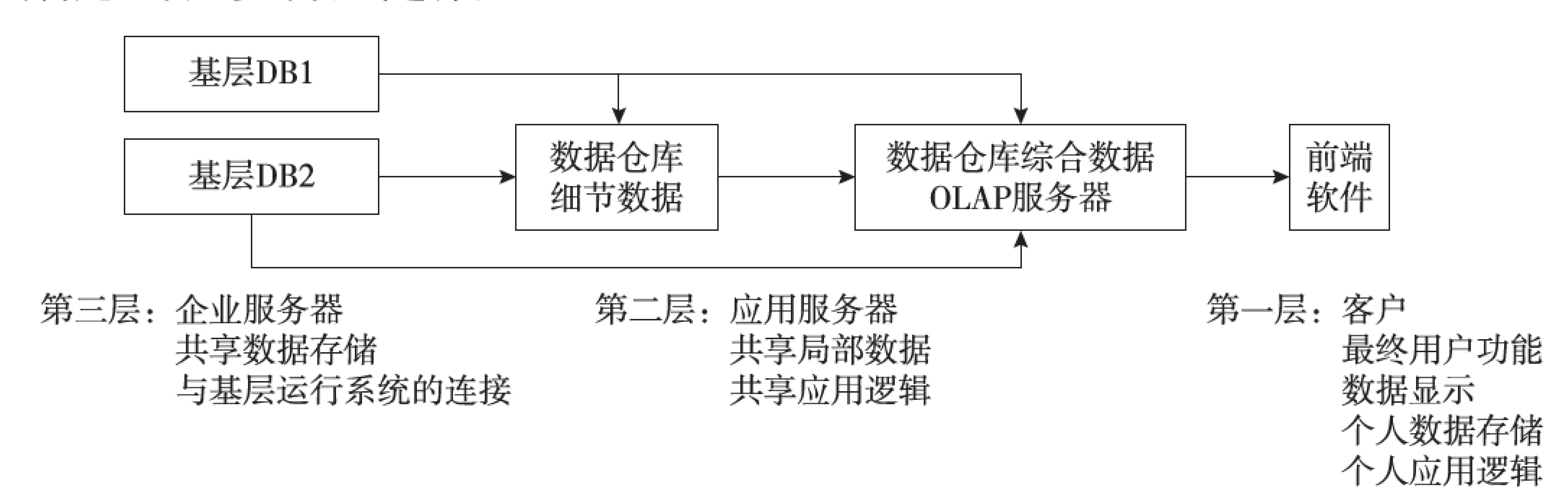


图2.16　OLAP的三层客户/服务器逻辑结构图

显然，OLAP服务器的构建基础是数据仓库或基层数据库，而OLAP的对象是面向分析和管理决策人员的。决策人员一般对综合性数据更为关注，使得在数据了解过程中视角能够更高层次以及更具体。因此，数据仓库中综合数据的组织以及前端用户的多维数据分析需求的满足成为OLAP服务器的设计重点。

市场中的多种OLAP软件工具和工具集都是以多维数据分析为目的，满足决策或多维环境的特殊的查询和报告需求，这是它们的追求。它们基本上是遵从三层结构的。

5. OLAP产品介绍及选择

按照数据存储格式，有三种类型的OLAP产品，即MOLAP、ROLAP和HOLAP①。

其中，MOLAP产品主要有Cognos的Powerplay、Hyperion的Essbase和微软的Analysis Service等。这类产品从关系数据库（甚至是文本文件、Excel文件）中抽取数据，并存储在自己的数据库中。与Oracle、DB2这类关系数据库不同，这种数据库并没有标准的访问接口，而是基于专有格式。因此，这些产品实现多维存储的原理也不尽相同，但其数据的存放大致是以编程语言中多维数组的方式为主。数组的每个维对应一个维度，数组的单元格中存放度量值。维度与维元素的数量在极大程度上影响多维数据库的单元格数量，因此，数据库随着维度的增加也迅速膨胀，对于MOLAP产品来说，多维存储的存储空间与性能就显得极为关键。Essbase在这方面做了很多优化工作，但有时也会显得过于复杂，Powerplay则采用了比较简单的优化方法，提供某些选项（如cube分区等）。

提供多维存储是OLAP产品的核心功能，除此之外，它们也能够将用户通过前端发出的OLAP访问操作转换为对数据的请求并予以返回，因此，就需要考虑有哪些前端工具能够与OLAP产品对接。

Cognos的Powerplay是个相对封闭的产品，不能用其他前端来访问，它有自己的客户端和Web Explorer。而Hyperion和微软则与Powerplay不同，采用的是开放式接口，第三方可以利用

① http://www.doc88.com/p-614824598430.html.

它们提供的丰富的API访问其数据库。事实上，一些第三方的前端工具与OLAP产品的对接正是基于微软开发的MDX和参与的XMLA（XML for Analysis）规范实现的，比如利用BO WebI连接Essbase。微软的服务器甚至像用SQL来访问关系数据库一样，提供了MDX来查询多维数据。

由于ROLAP产品的数据是存放在关系数据库中的，因此它对关系模型的要求就显得十分严格，比如为了定义像维度、度量、事实表、聚集表等元数据，就必须遵循星型模式或雪花模式。但这样就使得部署的难度增加了，并且如果聚集表构建得不好，很难保证最后的访问性能。MicroStrategy就是ROLAP产品。

目前，也有很多混合型OLAP产品，这是因为，将一些ROLAP的特性增加到现有数据库上，对那些本身就做关系数据库的厂商来说，并不是一件难事。在与Essbase终止OEM合同之后，IBM推出一个名为CubeViews的产品，可以说这就是一个ROLAP产品。

OLAP服务器和工具可以根据5个方面来进行评价：特征和功能、访问性能、OLAP服务引擎、管理以及全局结构。用户可以根据这5个方面分析市场上的OLAP产品，也可以把它们作为应用系统中的OLAP需求分析指标。

2.5 练习

1. 数据存储的主要模式有哪些?
2. 新兴数据存储系统与传统关系型数据存储有哪些不同?
3. 海量数据存储的关键技术有哪些?
4. 数据仓库的定义和特点分别是什么?
5. 数据仓库与数据库有什么不同?
6. 数据集市的定义是什么？有什么特点?
7. 什么是元数据?
8. 数据仓库设计的主要步骤有哪些?
9. ETL处理的主要过程是什么?
10. 数据仓库存储的数据模型有哪些？说明它们的不同点。
11. OLAP技术用于数据仓库时，如何提高数据仓库的分析能力。

参考文献

[1] Viktor Mayer–Sch. nberger, Kenneth Cukier. 大数据时代[M]. 杭州：浙江人民出版社，2012.
[2] Paulraj Ponniah. 数据仓库基础[M]. 北京：电子工业出版社，2004.
[3] 鲍亮，李倩. 实战大数据[M]. 北京：清华大学出版社，2014.
[4] willian H.Tnmon. 数据仓库[M]. 北京：机械工业出版社，2006.
[5] 何玉洁，张俊超. 数据仓库与OLAP实践教程[M]. 北京：清华大学出版社，2008.
[6] 陈文伟，黄金才. 数据仓库与数据挖掘[M]. 北京：人民邮电出版社，2004.

[7] Ralph Kimball, Margy Ross. 数据仓库工具箱：维度建模的完全指南[M]. 北京：电子工业出版社，2003.

[8] 殷利国. 电视台制播系统中存储方案的设计[N]. 广播与电视技术，2007，3：94–96.

[9] 魏薇. 一种基于嵌入式Linux的NAS模型实现[D]. 电子科技大学硕士论文，2004：4–13.

[10] 杨丰滔. 基于SAN的存储资源管理系统（SPM）的研究与开发[D]. 西北工业大学硕士论文，2003：10–12.

[11] 李延光. 基于Hadoop的海量工程数据处理技术研究[D]. 北京交通大学硕士论文，2013：9–12.

[12] 拓守恒. 云计算与云数据存储技术研究[J]. 电脑开发与应用，2010，23（9）：1–2.

[13] 王武龙. 基于遗传算法的聚类数据挖掘及其在销售系统中的应用[D]. 大连：大连交通大学电气信息学院，2002.

[14] 孙吉赟. 数据仓库多结构粒度模型与计算机研究[D]. 西安：西安石油大学，2008.

[15] 钟静华. 数据仓库中物化视图选择算法的研究[D]. 厦门：厦门大学，2006.

[16] 王珊. 数据仓库技术与联机分析处理[M]. 北京：科学出版社，1999.

[17] 张城. 商业智能系统在零售行业中的研究与应用[D]. 青岛科技大学硕士论文，2008：18–21.

[18] 毕然. 一个基于实视图的ROLAP系统的设计与实现[D]. 东南大学硕士论文，2004：43–44.

[19] 孙瑞超. 基于J2EE和设计模式的DSS平台研究与应用[D]. 大连理工大学硕士论文，2005：10–11.

[20] 杨静. 数据仓库技术在高校科研管理中的应用研究[D]. 河北工程大学硕士论文，2008：20.

第3章 NoSQL

NoSQL，泛指非关系型数据库。相对于传统关系型数据库，NoSQL有着更复杂的分类：key-value数据库、文档数据库、Column-oriented数据库以及图存数据库等。这些类型的数据库能够更好地适应复杂类型的海量数据的存储。本章介绍了NoSQL的相关概念、应用现状以及数据一致性理论等内容，并对key-value数据库、Column-oriented数据库、图存数据库、文档数据库、NewSQL数据库以及分布式缓存系统等内容作了详细的介绍。

3.1 NoSQL简介

本节对NoSQL的概念、发展以及应用现状等内容进行了详细的介绍，并结合传统的关系型数据库，分析了NoSQL数据库的特点。

3.1.1 什么是NoSQL

1998年，Carlo Strozzi提出NoSQL一词，用来指代他所开发的一个没有提供SQL功能的轻量级关系型数据库。顾名思义，此时的NoSQL可以被认为是“No SQL”的合成。

2009年初，Johan Oskarsson发起了一场关于开源分布式数据库的讨论，Eric Evans在这次讨论中再次提出了NoSQL的概念。此时，NoSQL主要指代那些非关系型的、分布式的且可不遵循ACID原则的数据存储系统。这里的ACID是指Atomic（原子性）、Consistency（一致性）、Isolation（隔离性）和Durability（持久性）。

同年，在亚特兰大举行的“no：sql（east）”讨论会，无疑又推进了NoSQL的发展。此时，它的含义已经不仅仅是“No SQL”这么简单，而演变成了“Not Only SQL”，即“不仅仅是SQL”。因此，NoSQL具有了新的意义：NoSQL数据库既可以是关系型数据库，也可以是非关系型数据库，它可以根据需要选择更加适用的数据存储类型。

NoSQL的整体框架如图3.1所示。

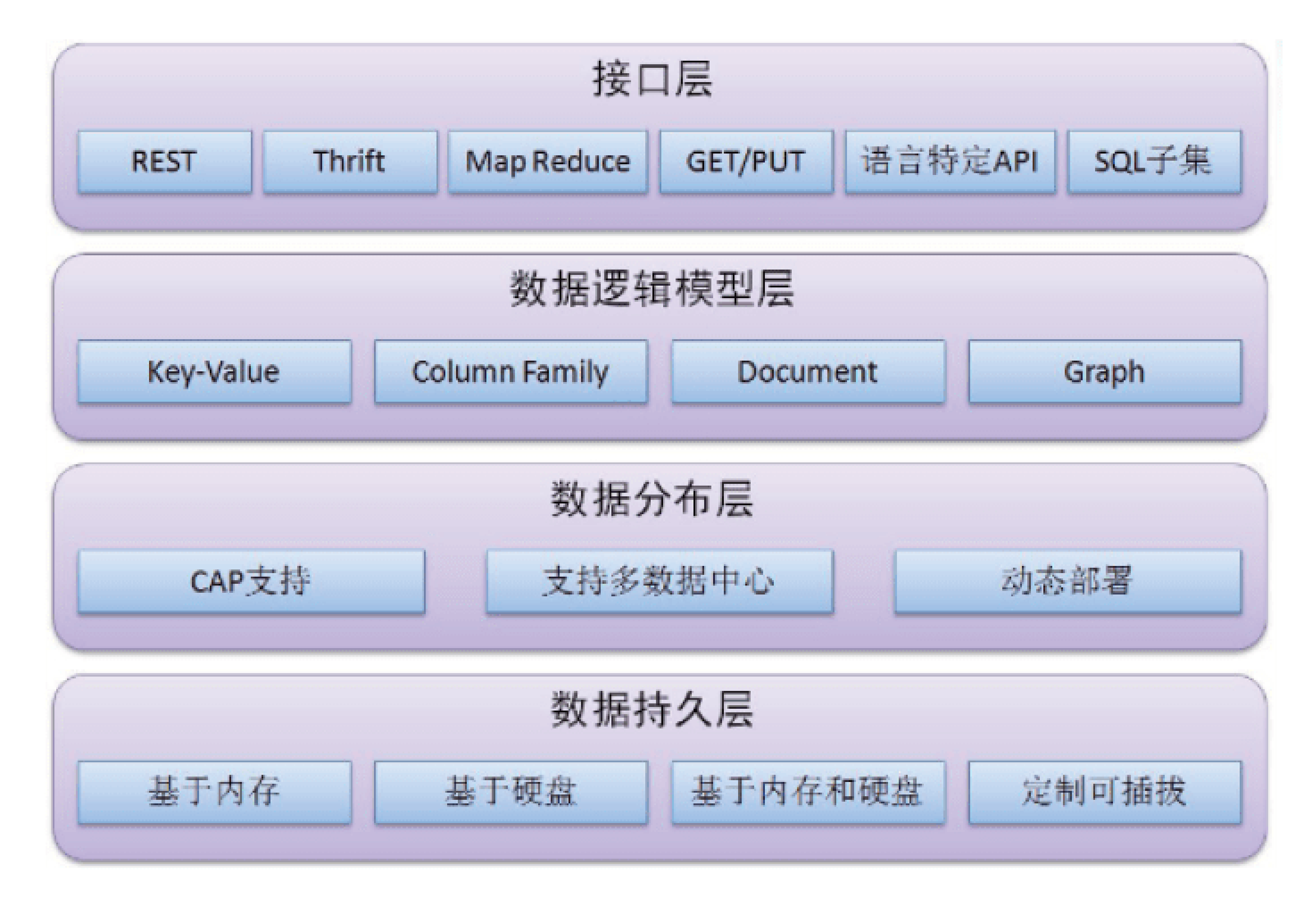


图3.1　NoSQL的整体框架

典型的NoSQL数据库主要分为：Key-Value数据库、Column-oriented数据库、图存数据库和文档数据库，如图3.2所示。对这些数据库的具体介绍在3.3~3.6节。

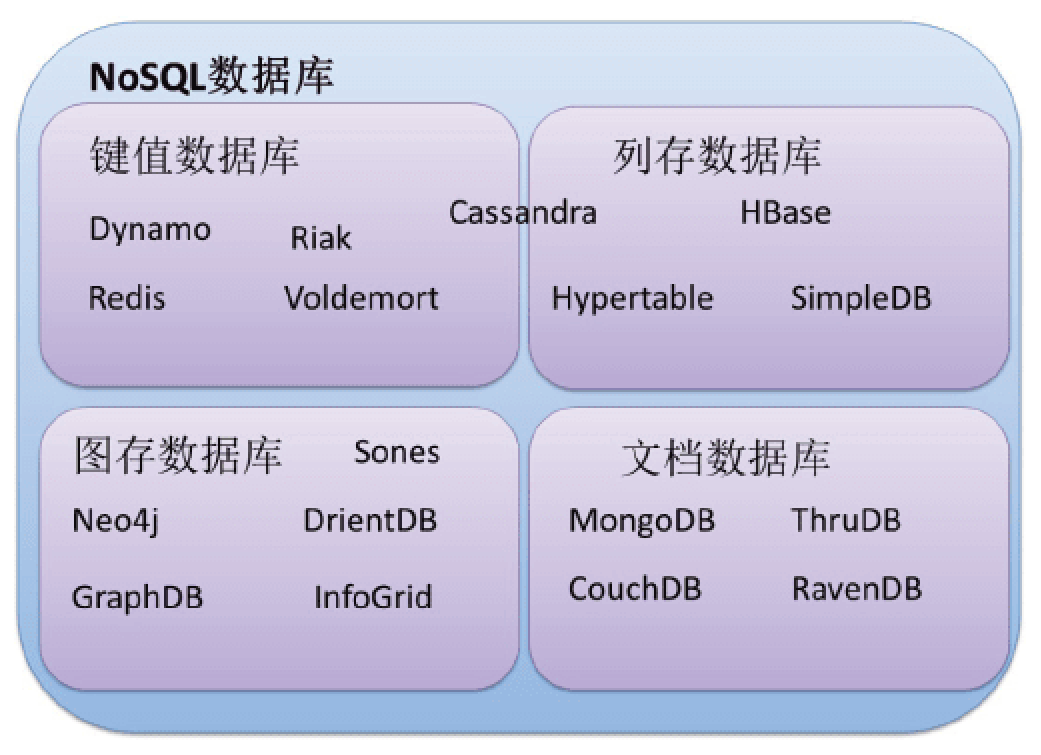


图3.2　典型的NoSQL分类

（1）key-value数据库

key-value存储是最常见的NoSQL数据库存储形式。key-value数据库存储的优势是处理速度非常快，它的缺点是只能通过键的完全一致查询来获取数据。根据数据的保存方式，可分为临时性、永久性和两者兼具三类。

临时性键值存储是在内存中保存数据，可以进行非常快速的保存和读取处理，数据有可能丢失，比如memcached。永久性键值存储是在硬盘上保存数据，可以进行非常快速的保存和读取处理，虽然无法与memcached相比，但数据不会丢失，比如Tokyo Tyrant、ROMA等。两者兼具的键值存储可以同时在内存和硬盘上保存数据，进行非常快的保存和读取处理，并且保存在硬盘上的数据不会消失，即使消失也可以恢复，适合于处理数组类型的数据，比如Redis。

（2）Column-oriented数据库

普通的关系型数据库都是以行为单位来存储数据的，擅长进行以行为单位的读入处理。而NoSQL的列数据库是以列为单位来存储数据的，因此擅长以列为单位来读取数据。行数据库可以对少量行进行读取和更新，而列数据库可以对大量行少量列进行读取，同时对所有行的特定列进行更新。Column-oriented数据库具有高扩展性，即使增加数据也不会降低相应的处理速度。主要产品有Bigtable，Apache Cassandra等。

（3）图存数据库

图存数据库主要是指将数据以图的方式存储。实体被作为顶点，实体之间的关系则被做为边。比如有三个实体，Steve Jobs、Apple和Next，会有两个“Founded by”的边将Apple和Next连接到Steve Jobs。图存数据库主要适用于关系较强的数据，但适用范围很小，因为很少

有操作涉及到整个图。主要产品如Neo4j、GraphDB、OrientDB等。

（4）文档数据库

文档数据库是一种用来管理文档的数据库，它与传统数据库的本质区别在于，其信息处理基本单位是文档，可长、可短、甚至可以无结构。在传统数据库中，信息是可以被分割的离散数据段。文档数据库与文件系统的主要区别在于文档数据库可以共享相同的数据，而文件系统不能，同时，文件系统比文档数据库的数据冗余复杂，会占用更多的存储空间，更难于管理维护。文档数据库与关系数据库的主要区别在于，文档数据库允许建立不同类型的非结构化或者任意格式的字段，并且不提供完整性支持。但是它与关系型数据库并不是相互排斥的，它们之间可以相互补充、扩展。文档数据库的两个典型代表是CouchDB和MongoDB。

3.1.2 什么是关系型数据库

1969年，Edgar Frank Codd发表一篇跨时代的论文，首次提出了关系数据模型的概念。但由于论文“IBM Research Report” 只是刊登在IBM公司的内部刊物上，所以反响平平。1970年，他再次发表了题为“A Relational Model of Data for Large Shared Data banks”的论文并刊登在《Communication of the ACM》上，才引起了大家的关注。

现今关系型数据库的基础就是采用由Codd提出的关系数据模型。由于当时的硬件性能低劣、处理速度过慢，关系型数据库迟迟没有得到实际应用。随着硬件性能的提升，加之具有使用简单、性能优越等优点，关系型数据库才得到了广泛应用。

关系型数据库是建立在关系模型基础上的数据库，借助于集合代数等数学概念和方法来处理数据库中的数据。即把所有的数据都通过行和列的二元表现形式表示出来，给人更容易理解的直观感受。现实世界中的各种实体以及实体之间的各种联系均可表示为关系模型。

经过数十年的发展，关系型数据库已经变得比较成熟，目前市场上主流的数据库都为关系型数据库，比较知名的有：Sybase，Oracle，Informix，SQL Server和DB2等。

3.1.3 NoSQL数据库与关系型数据库的比较

NoSQL数据库和传统的关系型数据库都具备各自的特点，本节从优势与缺陷、应用现状等方面分析了两种类型数据库的特点。

1. 关系型数据库的优势

关系型数据库相比于其他模型的数据库，有以下几点优势。

（1）容易理解。相对于网状、层次等其他模型来说，关系模型中的二维表结构非常贴近逻辑世界，更容易理解。

（2）便于维护。由于丰富的完整性，使数据冗余和数据不一致的概率大大降低。

（3）使用方便。操作关系型数据库时，只需使用SQL语言在逻辑层面进行操作即可。

2. 关系型数据库存在的问题

传统的关系型数据库具有高稳定型、操作简单、功能强大、性能良好的特点，同时也积累了大量成功的应用案例。上世纪90年代的互联网领域，一个网站的访问量用单个数据库就

已经足够，而且当时静态网页占绝大多数，动态交互类型的网站相对较少。

随着互联网中Web 2.0网站的快速发展，微博、论坛、微信等逐渐成为引领Web领域的潮流主角。在应对这些超大规模和高并发的纯动态网站时，传统的关系型数据库就遇到了很多难以克服的问题。同时，根据用户个性化信息，高并发的纯动态网站一般可以实时生成动态页面和提供动态信息。鉴于这种数据库高并发读写的特点，它基本上无法使用动态页面的静态化技术，因此数据库并发负载往往会非常高，一般会达到每秒上万次的读写请求。然而关系数据库只能应付上万次SQL查询，面对上万次的SQL写数据请求，硬盘的输入/输出端就显得无能为力了。

此外，在以下两方面，关系型数据库也存在问题。海量数据的高效率存储及访问：对于关系型数据库来说，Web 2.0网站的用户每天都会产生海量的动态信息，因此在对一张数以亿计的记录表进行SQL查询，效率是极其低下的。数据库的高可用性和高可扩展性：由于Web架构的限制，数据库无法再添加硬件和服务节点来扩展性能和负载能力，尤其对需要提供24小时不间断服务的网站来说，数据库系统的升级和扩展只能通过停机来实现，这样的决定将会带来巨大的损失。

3. NoSQL数据库的优势

虽然NoSQL只应用在一些特定的领域上，但它足以弥补关系型数据库的缺陷。NoSQL的优势主要有以下四点。

（1）NoSQL比关系型数据库更容易扩散。虽然NoSQL数据库种类繁多，但由于它们能够去掉关系型数据库的关系特性，从而使得数据之间无关系，这样就非常容易扩展，进而为架构层面带来了可扩展性。

（2）NoSQL比一般数据库具有更大的数据量，而且性能更高。这主要得益于它的无关系性，数据库的结构简单。比如在针对Web 2.0的交互频繁地应用时，由于MySQL的Cache是大粒度的，性能不高，故MySQL使用Query Cache时，每次表更新Cache就失效；然而NoSQL里的Cache是记录级的，是一种细粒度的Cache，所以就这个层面来说，NoSQL的性能就高很多了。

（3）NoSQL具有灵活的数据模型。NoSQL不需要事先为要存储的数据建立字段，它可以随时存储自定义的数据格式。而在关系型数据库里，增删字段却是一件非常麻烦的事情。这一点在Web 2.0大数据时代更为明显。

（4）NoSQL的高可用性。在不太影响其他性能的情况下，NoSQL也可以轻松地实现高可用的架构。如Cassandra模型和HBase模型，就可以通过复制模型来实现高可用性。

4. NoSQL数据库的实际应用缺陷

（1）缺乏强有力的商业支持。目前NoSQL数据库绝大多数是开源项目，没有权威的数据库厂商提供完整的服务，在使用NoSQL产品时，如果出现故障，就只能依靠自己解决，因此在这方面需要承担较大的风险。

（2）成熟度不高。NoSQL数据库在现实当中的实际应用较少， NoSQL的产品在企业中也并未得到广泛的应用。

（3）NoSQL数据库难以体现实际情况。由于NoSQL数据库不存在与关系型数据库中的关系模型一样的模型，因此对数据库的设计难以体现业务的实际情况，这也就增加了数据库设

计与维护的难度。

5. NoSQL数据库应用现状

NoSQL数据库存在了十多年，有很多成功案例，受欢迎程度近期更是在不断增加，其中原因主要有以下两个方面。

（1）随着社会化网络和云计算的发展，以前只在高端组织才会遇到的一些问题，现在已经普遍存在了。

（2）现有的方法随着需求一起扩展，并且很多组织不得不考虑成本的增加，这就要求它们去寻找性价比更高的方案。

6. 关系型数据库与NoSQL数据库结合

分布式存储系统更适合用NoSQL数据库，现有的Web 2.0网站会遇到的问题也会迎刃而解。但是NoSQL数据库在实际应用上的缺陷又让用户难以放心。这使很多开发人员考虑将关系型数据库与NoSQL数据库相结合，在强一致性和高可用性场景下，采用ACID模型；而在高可用性和扩展性场景下，采用BASE模型。虽然NoSQL数据库可以对关系型数据库在性能和扩展性上进行弥补，但目前NoSQL数据库还难以取代关系型数据库，所以才需要把关系型数据库和NoSQL数据库结合起来使用，各取所长。

图3.3为数据库的系统分类，从中可更好地了解它们之间的关系。

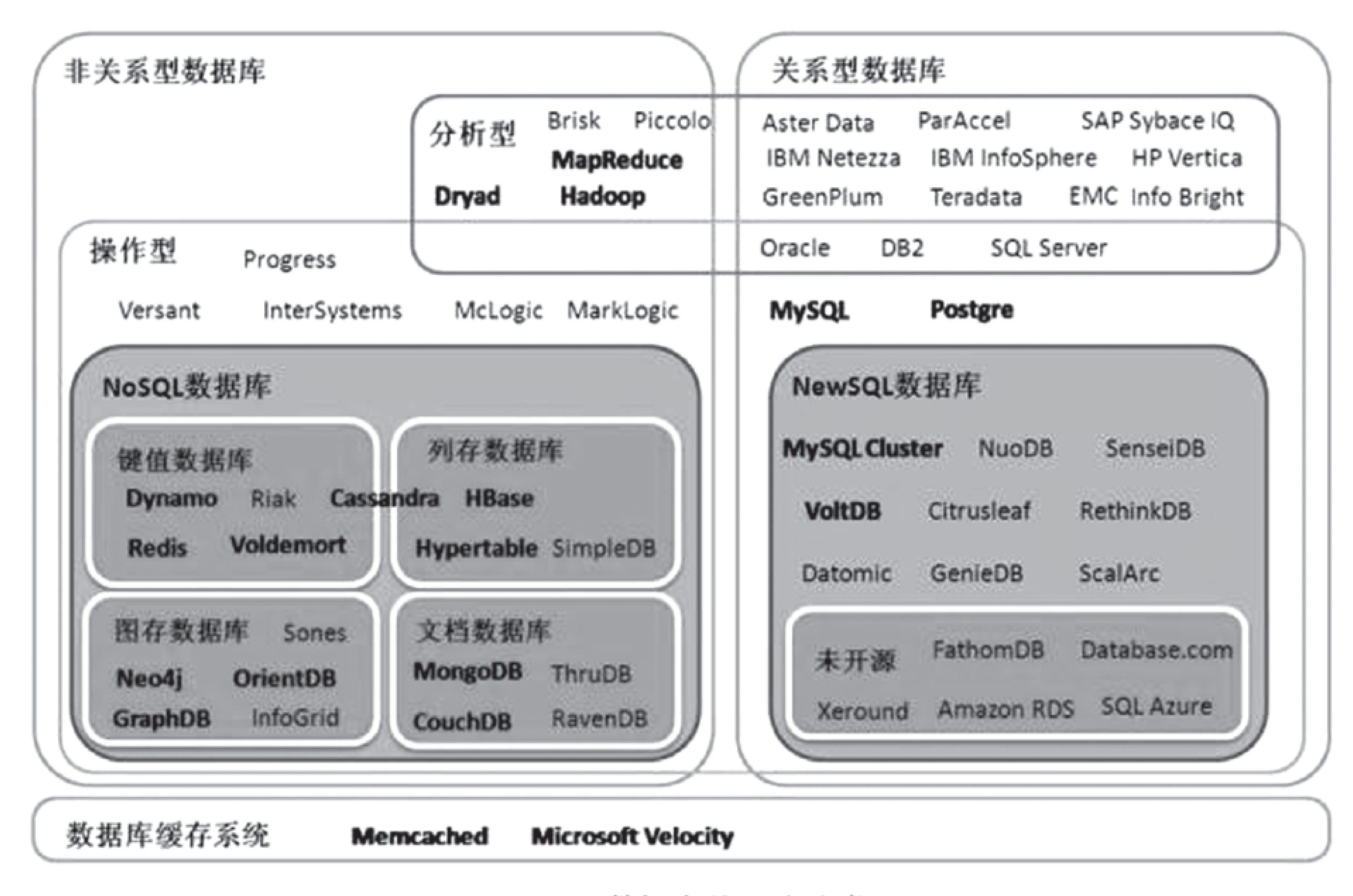


图3.3 数据库的系统分类

3.2 NoSQL的三大基石

从NoSQL的优势来看，NoSQL的优势主要得益于它在海量数据管理方面的高性能。而海

量数据管理所涉及到的存储放置策略、一致性策略、计算方法、索引技术等都是建立在数据一致性理论的基础之上的。数据一致性理论又包括：CAP理论、BASE模型和最终一致性，是NoSQL的三大基石。本节会对数据一致性理论进行详细地介绍。

3.2.1 CAP

2000年，Eric Brewer在ACM PODC会议中提出了CAP理论，该理论又被称为BrewerL理论。“C”、“A”、“P”分别代表一致性（Consistency）、可用性（Avaliability）、分隔容忍性（Partition tolerance），如图3.4所示。

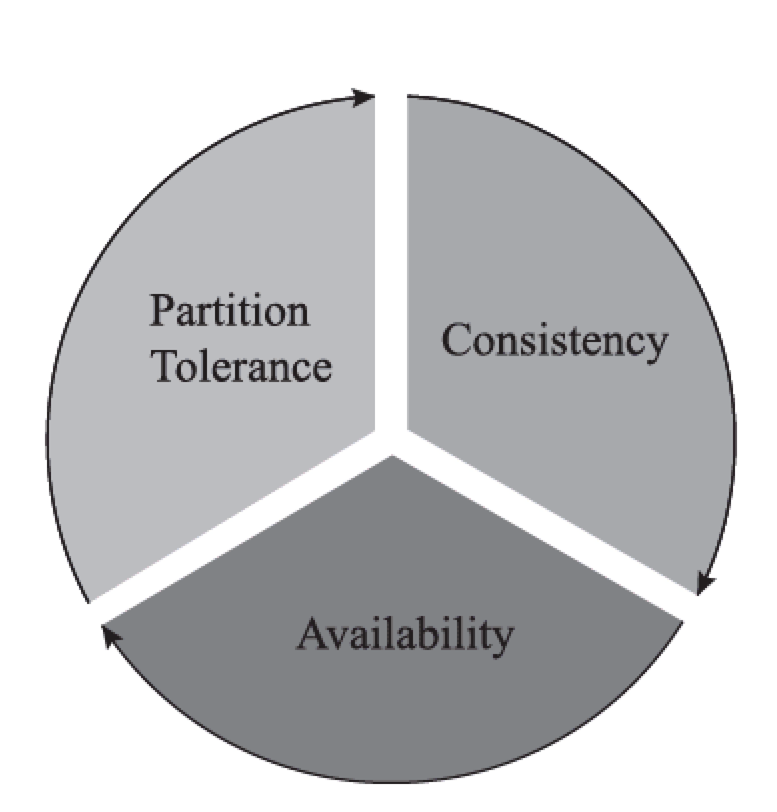


图3.4 CAP理论

1. 一致性（Consistency）

在分布式计算系统中，在执行过某项操作之后，所有节点仍具有相同的数据，这样的系统被认为具有一致性。

2. 可用性（Avaliability）

在每一个操作之后，无论操作成功或者失败都会在一定时间内返回相应结果。下面将对一定时间内和返回结果进行详细解释。

一定时间内是指系统操作之后的结果应该在给定的时间内反馈。如果超时则被认为不可用，或者操作失败。比如进入系统时进行账号登录，在输入相应的登录密码之后，如果等待时间过长，如3分钟，系统还没有反馈登录结果，登录者将会一直处于等待状态，无法进行其他操作。

返回结果也是很重要的因素。假如在登录系统之后，结果是出现“java.lang.error……”之类的错误信息，这对于登录者来说相当于没有返回结果。他无法判断自己登录的状态，是成功还是失败，或者需要重新操作。

3. 分隔容忍性（Partition tolerance）

分隔容忍性可以理解为在网络由于某种原因被分隔成若干个孤立的区域，且区域之间互不相同时，仍然可以接受请求。当然也有一些人将其理解为系统中任意信息的丢失或失败都不会影响系统的继续运作。

CAP理论指出，在分布式环境下设计和部署系统时，只能满足上面三个特性中的两项，不能满足全部三项。所以，设计者必须在三个特性之间做出选择。

然而，分布式系统为什么不能同时满足CAP理论的三个特性呢?

在正常情况下，系统的操作步骤如下：

（1）A将V_0更新为数据值V_1。

（2）G_1将消息m发送给G_2，G_1中的数据V_0更新为V_1。

（3）B读取到G_2中的数据V_1。

上述步骤可用图3.5表示。

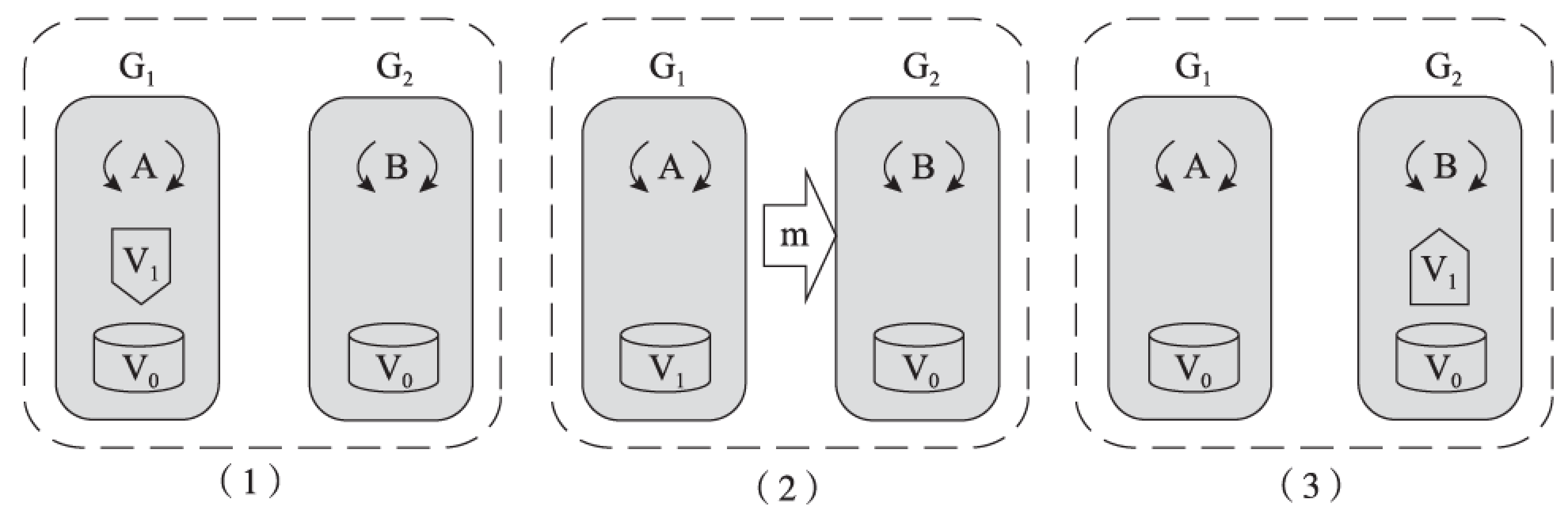


图3.5　正常情况

假设G_1和G_2分别代表网络中的两个节点，V_0是两个节点上存储的同一数据的不同副本，A和B分别是G_1和G_2上与数据交互的应用程序。

如果步骤（2）发生错误，即G_1的消息不能发送给G_2，此时B读取到的就不是更新的数据V_1，这样就无法满足一致性C。如果采用一些技术，如阻塞、加锁、集中控制等来保证数据的一致性，那么必然会影响到可用性A和分隔容忍性P。即使对步骤（2）加上一个同步消息，尽管能够保证B读取到数据V_1，但这个同步操作必定消耗一定的时间，尤其在节点规模成百上千的时候，不一定能保证可用性。也就是说，在同步的情况下，只能满足"C"和"P"，而不能保证"A"一定满足。

在如图3.6的例子中如果有一个事务组a，不妨假设为一组围绕着阻塞数据项V的工作单元，a1为写操作，a2为读操作。在一个local的系统中，可以利用数据库中的简单锁机制隔离a2中的读操作，直到a1的写成功完成。然而，在分布式的模型中，需要考虑到G_1和G_2节点，以及中间消息的同步可以完成。除非能够可以控制a2何时发生，否则永远无法保证a2可以读到a1写入的数据。所有加入阻塞、隔离、中央化的管理等控制方法，使影响分隔容忍性和a1（A）和a2（B）的可用性无法并存。

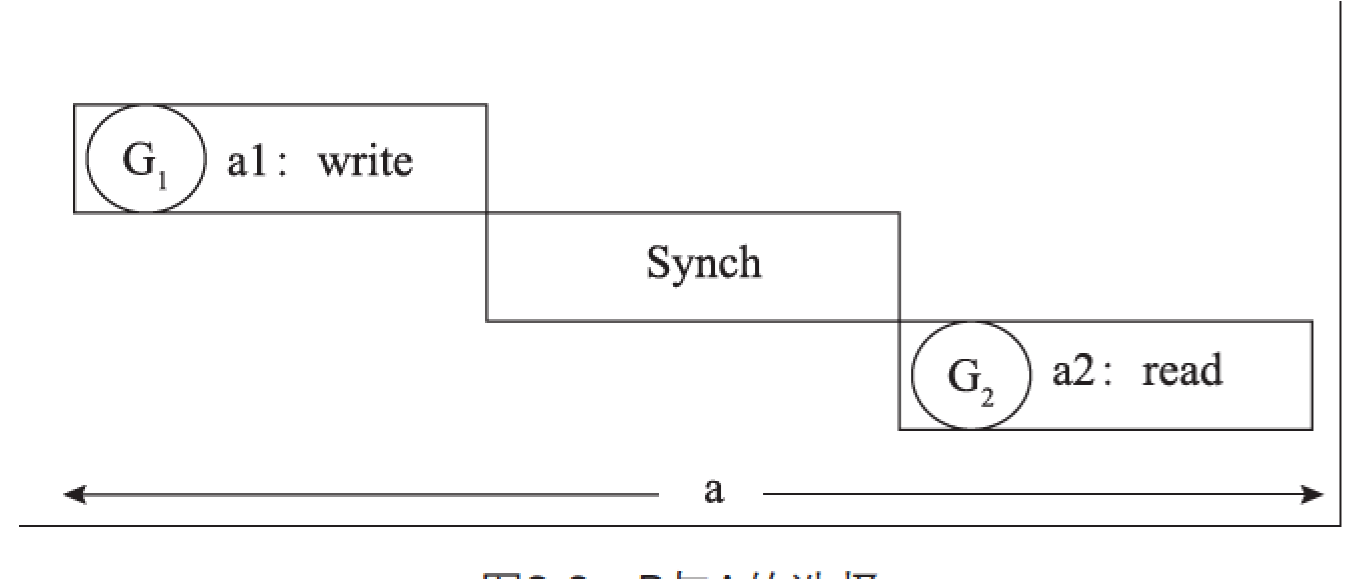


图3.6　P与A的选择

CAP中三种特性的不同选择如表3.1所示。

如果放弃P，即使将所有与事务有关的数据放到一台机器上，避免分隔带来的负面影响，也会严重影响系统的扩展性。

如果放弃A，一旦遇到分隔容忍故障，受影响的服务需要等待数据一致，并且在这个等待的时间段内，系统是无法对外提供服务的。

如果放弃C，这里放弃的一致性指的是放弃数据的强一致性，保留最终一致性。

表3.1　CAP问题的不同选择

序号	选择	特点	例子
1	C、A	两阶段提交、缓存验证协议	传统数据库、集群数据库、LDAP、GFS文件系统
2	C、P	悲观加锁	分布式数据库、分布式加锁
3	A、P	冲突处理、乐观	DNS、Coda

现在看来，如果理解CAP理论只是指多个数据副本之间读写一致性的问题，那么，它对关系数据库与NoSQL数据库来讲是完全一样的，它只是运行在分布式环境中的数据管理设施在设计读写一致性问题时需要遵循的一个原则而已，却并不是NoSQL数据库具有优秀的水平可扩展性的真正原因。如果将CAP理论中的一致性C理解为读写一致性、事务与关联操作的综合，则可以认为关系数据库选择了C与A，而NoSQL数据库则全都是选择了A与P，但并没有选择C与P的情况存在。也就是说传统关系型数据管理系统注重数据的强一致性，但是对于海量数据的分布式存储和处理，它的性能不能满足人们的需求，因此现在许多NoSQL数据库牺牲了强一致性来提高性能。CAP理论对于非关系型数据库的设计有很大的影响，这才是用CAP理论来支持NoSQL数据库设计的正确认识。这种认识正好与被广泛认同的NoSQL的另一个理论基础相吻合，即BASE。在3.2.2节中将会详细讲解BASE。

此时CAP的真正含义如图3.7所示。

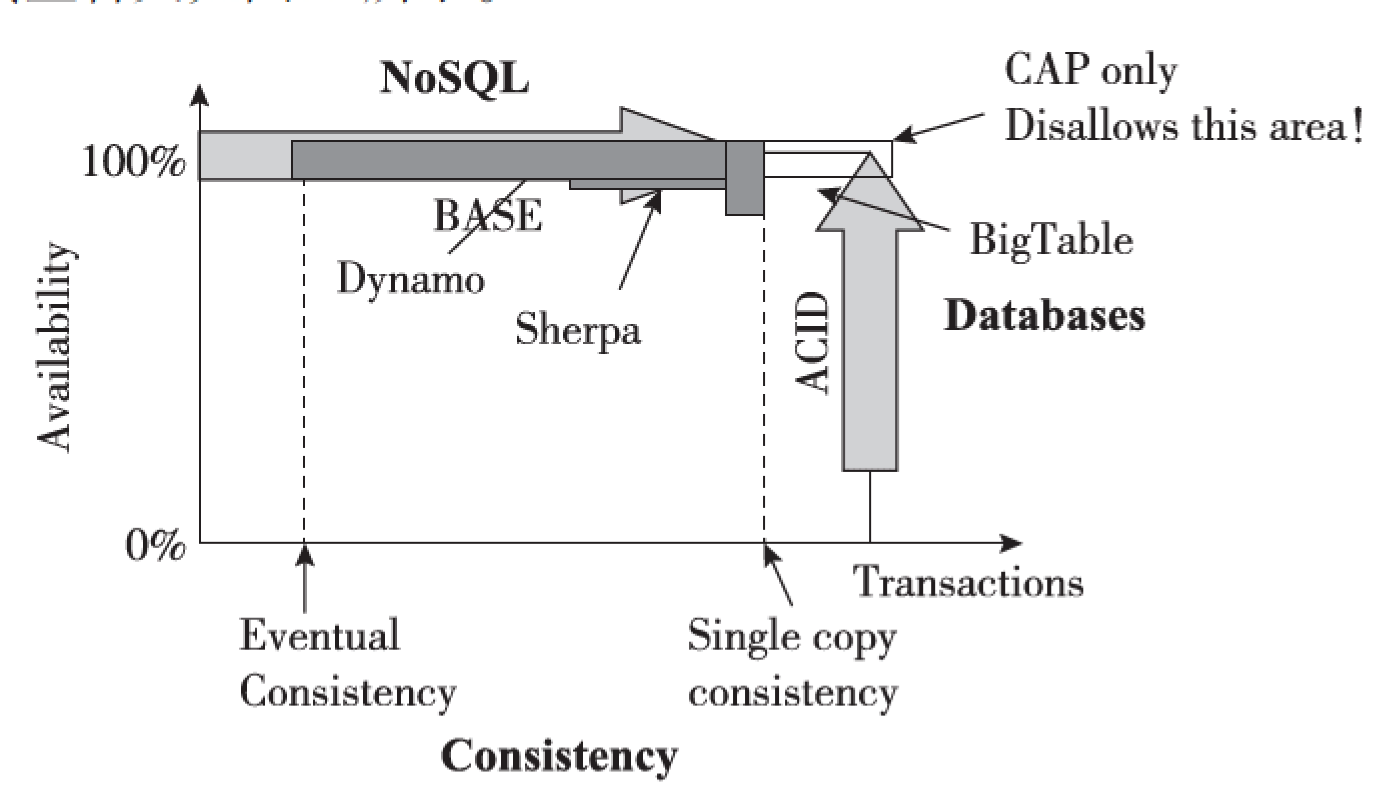


图3.7　CAP的真正含义

总之，CAP是为了探索适合不同应用的一致性C与可用性A之间的平衡。在没有发生分隔时，可以满足完整的C与A，以及完整的ACID事务支持。也可以通过牺牲一定的一致性C，来获得更好的性能与扩展性。在有分隔发生时，选择可用性A，集中关注分隔的恢复。需要分隔前、中、后期的处理策略，及合适的补偿处理机制。

3.2.2　BASE

BASE的含义是指NoSQL数据库设计可以通过牺牲一定的数据一致性与容忍性来换取高性能的保持甚至是提高，即NoSQL数据库都应该是牺牲C来换取P，而不是牺牲A，可用性A正好是所有NoSQL数据库都普遍追求的特性。BASE是缩写，说明如下。

- 基本可用（Basically Available）：系统能够基本运行、一直提供服务。
- 软状态（Soft-state）：系统不要求一直保持强一致状态。
- 最终一致性（Eventual consistency）：系统需要在某一时刻后达到一致性要求。

因此，BASE可以定义为CAP中AP的衍生。在单机环境下，ACID是数据的属性，而在分布式环境中，BASE就是数据的属性。BASE思想主要强调基本的可用性，即如果需要高可用性，也就是纯粹的高性能，那么就要以一致性或容忍性为牺牲。BASE的思想在性能方面还是有潜力可挖的。同时，BASE思想的主要实现有：按功能划分数据库和sharding碎片。

而且BASE的中文解释为碱，ACID的中文解释为酸，所以BASE与ACID是完全对立的两个模型。

ACID所代表的含义如下。

- 原子性（Atomicity）：事务中所有操作全部完成或者全部不完成。
- 一致性（Consistency）：事务开始或者结束时，数据库应该处于一致状态。
- 隔离性（Isolation）：假定只有事务它自己在操作数据库，且彼此之间并不知晓。
- 持续性（Durability）：一旦事务完成，就不能返回。

随着大数据时代的到来，系统数据（如社会计算数据，网络服务数据等）不断增长。对于数据不断增长的系统，它们对可用性及分隔容忍性的要求高于强一致性，并且很难满足事务所要求的ACID特性。而保证ACID特性是传统关系型数据库中事务管理的重要任务，也是恢复和并发控制的基本单位。

ACID与BASE的区别如表3.2所示。

表3.2　ACID与BASE的区别

ACID	BASE
强一致性	弱一致性
隔离性	可用性优先
采用悲观、保守方法	采用乐观方法
难以变化	适应变化、更简单、更快

3.2.3　最终一致性

在引入最终一致性之前先来介绍强一致性和弱一致性。

- 强一致性：无论更新操作是在哪个数据副本上执行的，之后的所有的读操作都会获得最新数据。
- 弱一致性：用户读到某一操作对系统特定数据的更新需要一段时间，这段时间被称为“不一致性窗口”。
- 最终一致性：是弱一致性的一种特例。在这种一致性系统下，保证用户最终能够读到某操作对系统特定数据的更新。

BASE是通过牺牲一定的数据一致性与容忍性来换取高性能的保持甚至提高。这里所说的牺牲一定的数据一致性并不是完全不管数据的一致性，否则数据将出现混乱，那么即使系统可用性再高、分布式再好也会没有任何利用价值。牺牲一致性，是指放弃关系型数据库中要

求的强一致性，只要系统能够达到最终一致性即可。

一致性可以从两个不同的视角来看，即客户端和服务端。从客户端角度来看，一致性指的是多并发访问时更新过的数据如何获取的问题。从服务端角度来看，一致性指的是更新如何复制分布到整个系统，以保证数据最终一致。一致性是因为有并发读写才出现的问题，因此在理解一致性的问题时，一定要结合考虑并发读写的场景。从客户端角度来看，多进程并发进行访问时，更新过的数据在不同进程如何获取不同策略，决定了不同的一致性。对于关系型数据库而言，要求更新过的数据都能被后续的访问看到，这是强一致性。如果能容忍后续的部分或者全部都访问不到，则就表现为弱一致性。如果要求一段时间后够能访问到更新后的数据，则为最终一致性。

根据更新数据后各进程访问到数据的方式和所花时间的不同，最终一致性模型又可以划分为以下五种模型。

- 因果一致性：假设存在A、B、C三个相互独立的进程，并对数据进行操作。如果进程A在更新数据将操作后通知进程B，那么进程B将读取A更新的数据，并一次写入，以保证最终结果的一致性。在遵守最终一致性规则条件下，系统不保证与进程A无因果关系的进程C一定能够读取该更新操作。
- “读己之所写”一致性：当某用户更新数据后，他自己总能够读取到更新后的数据，而且绝不会看到之前的数据。但是其他用户读取数据时，则不能保证能够读取到最新的数据。
- 会话一致性：这是“读己之所写”一致性模型的实用版本，它把读取存储系统的进程限制在一个会话范围之内。只要会话存在，系统就保证“读己之所写”一致性。也就是说，提交更新操作的用户在同一会话里读取数据时能够保证数据是最新的。
- 单调读一致性：如果用户已经读取某数值，那么任何后续操作都不会再返回到该数据之前的值。
- 单调写一致性：系统保证来自同一个进程的更新操作按时间顺序执行。这也叫做时间轴一致性。

以上五种最终一致性模型可以进行组合，例如“读己之所写”一致性与单调读一致性就可以组合实现，即读取自己更新的数据并且一旦读取到最新数据将不会再读取之前的数据。从实践的角度来看，这两者的组合，对于此架构上的程序开发来说，会减少额外的烦恼。

至于系统选择哪一种一致性模型，或者是哪种一致性模型的组合取决于应用对一致性的需求，而所选取的一致性模型会影响到系统处理用户请求及对副本维护技术的选择。

考虑系统一致性的需求，分布式存储在不同节点的数据将采用不同的数据一致性技术。例如：在关系型管理系统中一般会采用悲观方法（如加锁），而在一些强调性能的系统中则会采用乐观方法。

从服务端角度来看，如何尽快将更新后的数据分布到整个系统，降低达到最终一致性的时间窗口，是提高系统的可用度和用户体验非常重要的方面。这里主要讲解以下两种保证最终一致性的技术：类似于Quorum系统的一致性协议实现方法——Quorum系统的NRW策略；

两阶段提交协议。

1. Quorum系统的NRW策略

对于分布式数据系统，Quorum系统的一致性协议有三个关键值。

N：数据副本数。

W：写入数据时保证写完成所需的最小节点数。

R：读取数据时保证读完成所需的最小节点数。

如果W+R>N，即写的节点和读的节点存在重叠，则是强一致性。也就是说，在该策略中，只需要保证W+R>N，就可以保证强一致性。当W+R>N时，会产生与Quorum类似的效果。该模型中的读（写）延迟由最慢的R(W)副本决定。有时为了获得较高的性能和较小的延迟，R与W之和可能小于N，这时系统不能保证读操作能获取最新的数据。

例如：当N=3,W=2,R=2时，表示系统中数据副本数为3，在进行写操作时，需要等待至少两个副本完成了该写操作，系统才会返回执行成功的状态，对于读操作，系统有同样的特性。对于典型的一主一备同步复制的关系型数据库，当N=2,W=2,R=1时，则无论读的是主库还是备库的数据，都是强一致的。

对于典型的一主一备同步复制的关系型数据库，当N=2,W=1,R=1时，如果读的是备库，就可能无法读取主库已经更新过的数据，所以是弱一致性。

对于分布式系统，为了保证高可用性，一般设置N≥3。当R和W的值较小时，会影响一致性；当R和W的值较大时，会影响性能。因此，不同的N,W,R组合，是在可用性和一致性之间取一个平衡，以适应不同的应用场景。

以下是几种特殊场景：

当N=W,R=1时，系统对写操作的要求较高。任何一个写节点失效，都会导致写失败，同时可用性降低。但是由于数据分布的N个节点是同步写入的，因此可以保证强一致性。

当N=R,W=1时，系统对读操作较高。但只需要一个节点写入成功即可，写性能和可用性都比较高。若N个节点中有节点发生故障，那么读操作将不能完成。这种情况下，如果W<(N+1)÷2，并且写入的节点不重叠的话，则会存在写冲突。

当R=(N+1) ÷2，W=(N+1) ÷2时，系统兼顾了性能和可用性，在读与写之间取得平衡。Dynamo系统默认设置就是这种，即N=3,W=2,R=2。

2. 两阶段提交协议

两阶段提交协议（Two-phase Commitment Protocal——2PC协议）可以保证数据的强一致性。它把本地原子性提交行为的效果扩展到分布式事务，保证了分布式事务提交的原子性，并在不损坏日志的情况下，实现快速故障恢复，提高分布式数据库系统的可靠性。

在两阶段提交协议中，系统一般包含两类节点（或机器）：协调者（Coordinator）和参与者（Participants）。协调者一般情况下在一个系统里只有一个，而系统事务参与者则通常包含多个，这在数据存储系统中可以理解为多个数据副本。两类节点之间的关系框架如图3.8所示。

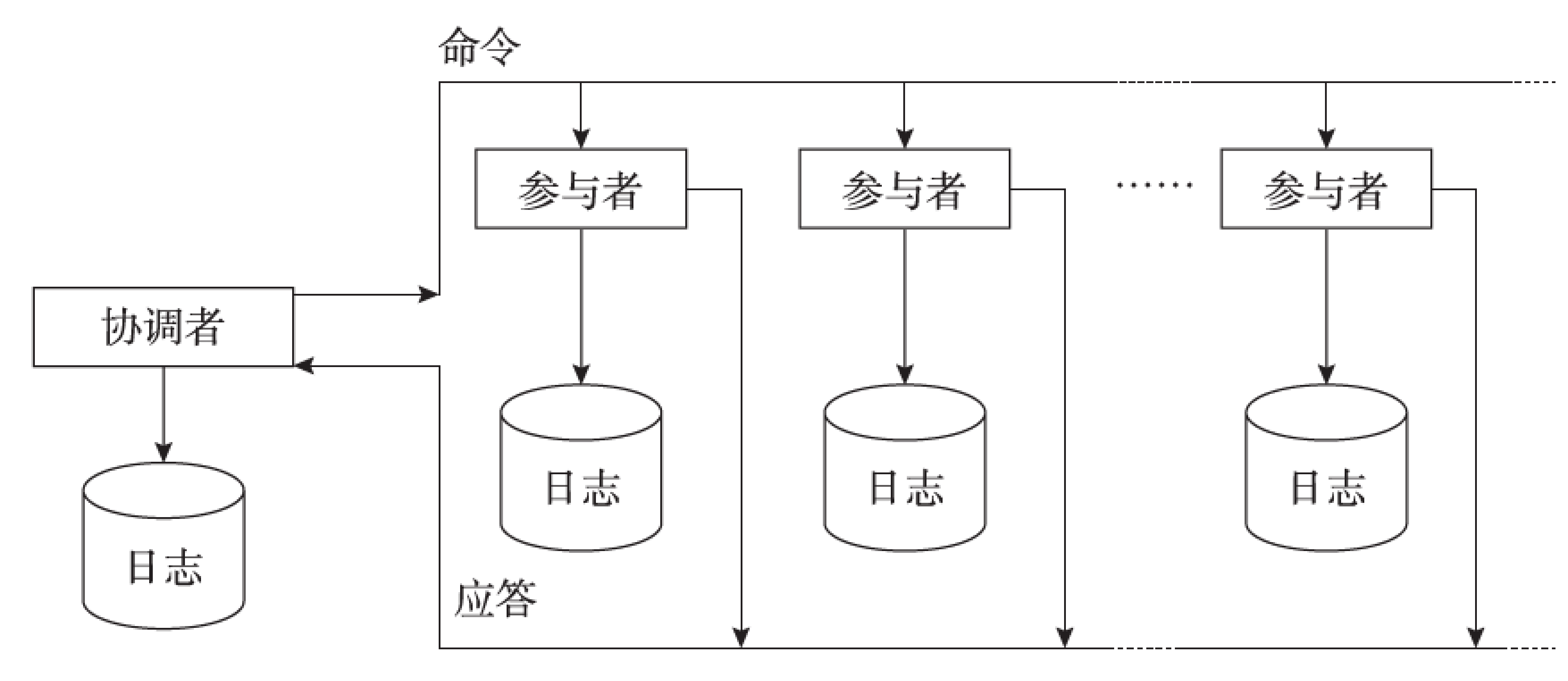


图3.8　协调者与参与者之间的关系框架

注意：一般情况下，只有协调者才有掌握提交或撤销事务的决定权，而其他参与者各自负责在其本地数据库中执行写操作，并向协调者提出撤销或提交子事务的意向。但是在两阶段提交协议中，允许参与者单方面撤销事务。一旦参与者确定了提交或撤销提议，则不能再更改它的提议，并且，当参与者处于就绪状态时，根据协调者发出的消息的种类，参与者可以转换为提交状态或撤销状态。其次，协调者依据全局提交规则做出全局终止决定。最后，注意协调者和参与者可能进入某些相互等待对方发送消息的状态。为了确保它们能够从这些状态中退出并终止，要使用定时器。每个进程进入一个状态时都要设置定时器。如果所期待的消息在定时器超时之前没有到来，定时器向进程报警，进程于是调用它自己的超时协议。

两阶段提交协议是由请求阶段和提交阶段两个阶段组成。图3.9、图3.10所示为两阶段提交协议活动。

阶段1：请求阶段（commit-request phase，或称表决阶段）

在请求阶段，协调者将通知事务参与者准备提交或取消事务，然后进入表决过程。在表决过程中，参与者将告知协调者自己的决策：同意或取消。同意的话，则事务参与者本地作业执行成功，取消的话，则事务参与者本地作业执行故障。

阶段2：提交阶段（commit phase，或称执行阶段）

在提交阶段，协调者将根据请求阶段的投票结果进行决策：提交或取消。当且仅当所有的参与者同意提交事务时，协调者才通知所有的参与者提交事务，否则协调者将通知所有的参与者取消事务。参与者在接收到协调者发来的消息后将执行相应的操作。

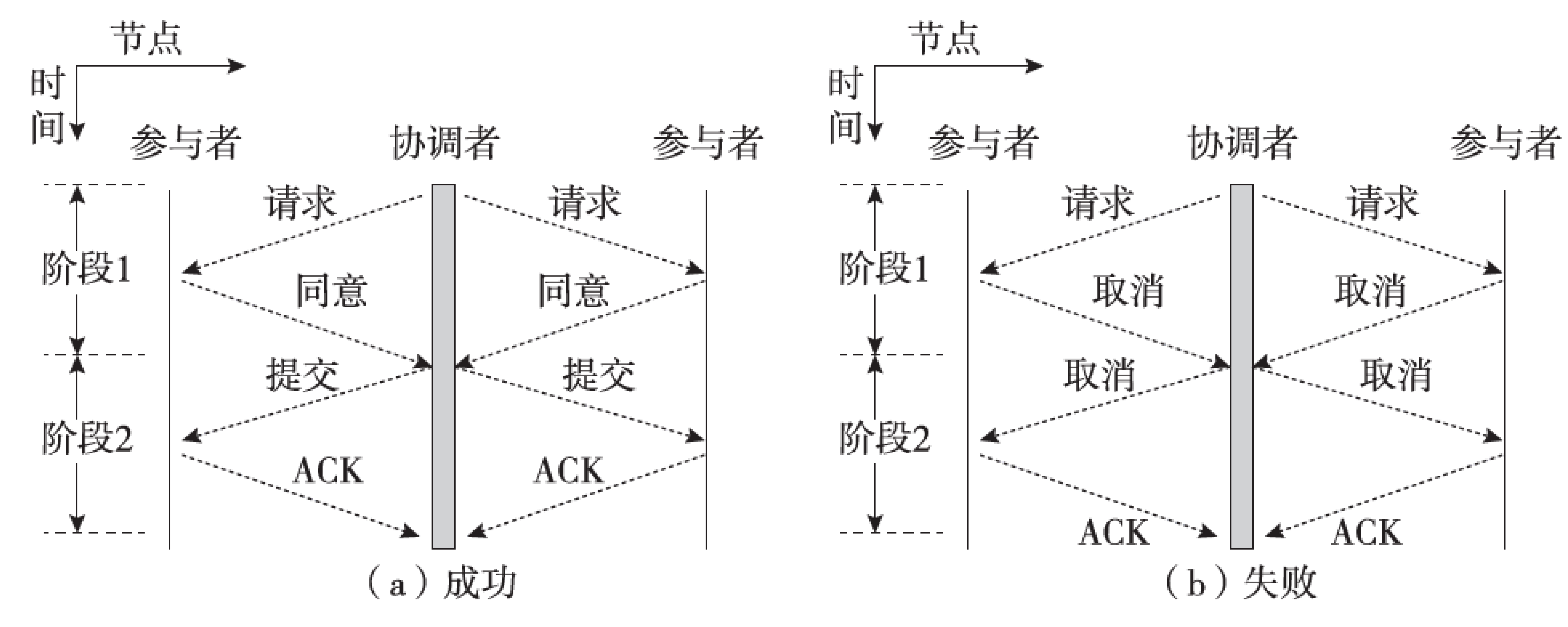


图3.9　两阶段提交协议1

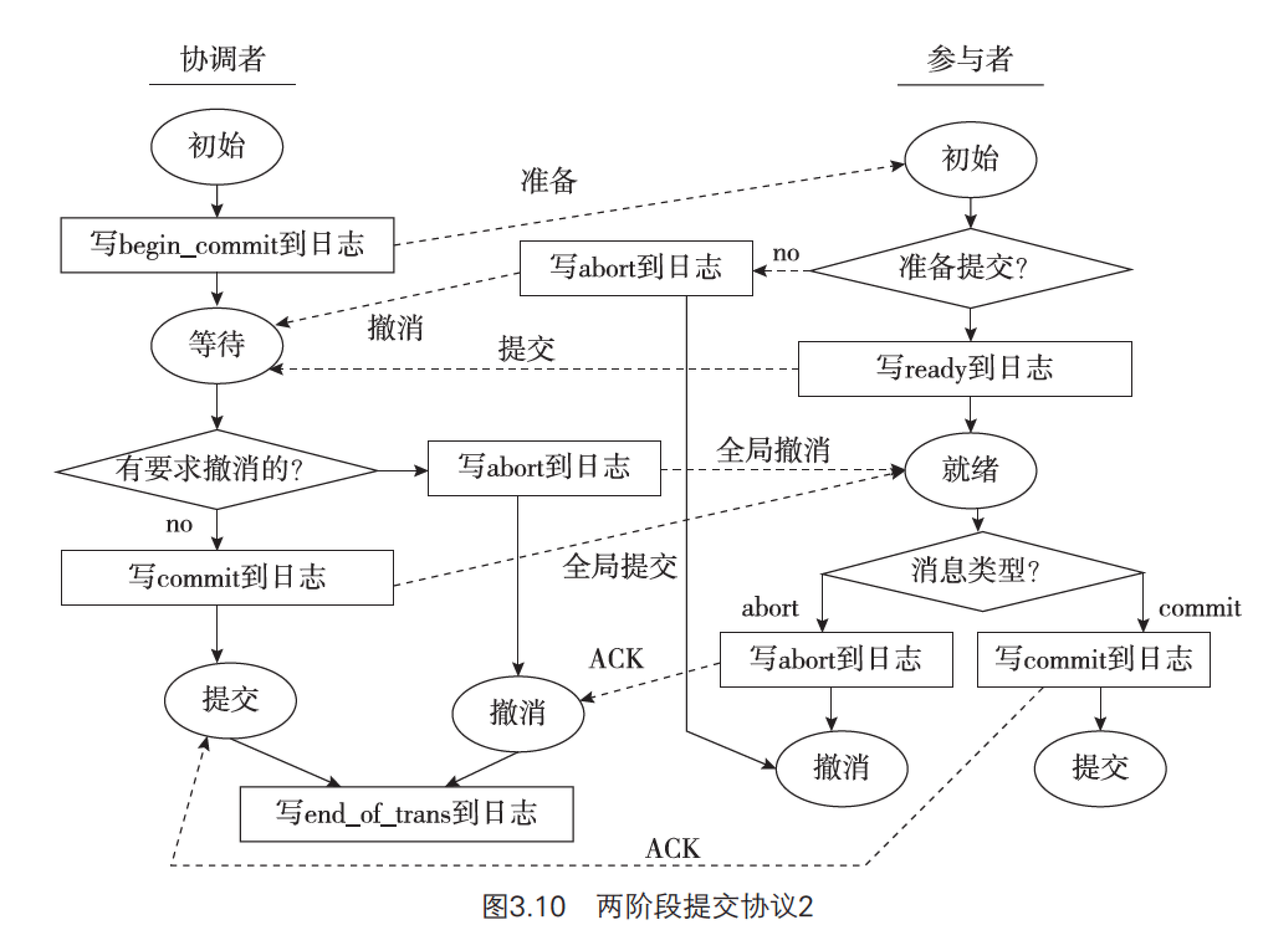


图3.10　两阶段提交协议2

图3.10描述了协调者和参与者之间的两阶段提交协议活动。这里参与者只有一个。图中椭圆形表示状态，虚线表示协调者和参与者之间的消息。虚线上的标号说明了消息的种类。

两阶段提交协议最大的缺点在于它是通过阻塞完成的协议，节点在等消息的时候处于阻塞状态，节点中其他进程也需要等待阻塞进程释放资源。如果协调者发生了故障，那么参与者将无法完成事务而一直等待下去。

如果参与者同意“提交消息”给协调者，但此时协调者发生永久故障，这样参与者将会一直等待，导致节点发生永久阻塞。同样，当协调者发送“请求提交”消息给参与者时，如果存在某个参与者发生永久故障，协调者将在某一时间内通知其他参与者取消该事务，不会一直阻塞。由于两阶段提交协议并没有容忍机制，因此如果一个节点发生故障，那么整个事务都将取消，而这也将付出较大的代价。

3.3　key-value数据库

键值存储是最常见的NoSQL数据库的存储形式。虽然它的处理速度非常快，但基本上只能通过键的完全一致查询获取数据。根据数据的保存方式，键值存储可分为临时性、永久性和两者兼具3种。这里将详细介绍两者兼具的键值存储——Redis。

3.3.1　Redis

Redis可以提供多种语言的应用程序编程接口（API），是一个开源的、可基于内存、支持网络、使用ANSI C语言编写的，可持久化的日志型key-value数据库。

与Memcached类似，Redis是一个key-value的存储系统，但它支持相对更多的value类型的存储，包括list（链表）、string（字符串）、set（集合）、zset（sorted set有序集合）和hash（哈希类型）。以上这些数据类型支持很多更具丰富性的原子性的操作，如add/remove、push/pop及取并集交集和差集等。Redis的数据都缓存在内存中，以便提高效率，与Memcached不同的是，Redis会周期性地把更新的数据写入到磁盘或把修改操作写入到追加的记录文件，并以此为基础实现master-slave（主从）同步①。

Redis支持主从同步。数据可以从主服务器向任意数量的从服务器上同步，从服务器可以是关联其他从服务器的主服务器，这使得Redis可执行单层树复制。从盘可以有意无意地对数据进行写操作。由于完全实现了发布/订阅机制，使得从数据库在任何地方同步树时，可订阅一个频道并接收主服务器完整的消息发布记录。同步对读取操作的可扩展性和数据冗余很有帮助。

Redis提供了五种数据类型：string、hash、list、set和zset（sorted set）。

1. string（字符串）

string（字符串）是Redis中最基本的数据类型，也是其余四种数据类型的基础，即它们都是由字符串类型组成的。一个Redis字符串可以包含任何种类型的数据，比如JPEG图像、序列化的Ruby对象。

2. Hash

Hash（哈希类型）是字符串字段和字符串值之间的映射，可以将Hashes类型看成具有String Key和String Value的map容器。该类型适合于存储值对象的信息，将一个对象存储在hash类型中比将对象的每个字段存成单个string类型占用的内存更少，而且对整个对象的存取操作更加简单方便。

3. List

Redis列表（list）是简单的字符串列表，实际上是使用双向链表的方式实现的。list（双向链表）的主要功能是pop、push、取得一个范围内的所有值等，list是一个链表结构，在操作中是把key理解为链表的名字。一般的操作是向列表两端添加、删除以及获取元素等。Redis列表访问元素的特点是：元素越接近列表的两端，获取该元素的速度越快；若通过索引访问元素，速度很慢，特别是在列表很长的情况下；若需要访问队列开头或结尾的前若干个元素，则访问数据的速度与队列的长度无关。

4. Set

Redis集合（set）是一个无序的字符串集合。不允许相同成员存在是Redis集合的特性，即向集合中添加相同的元素，只会存在一个元素，也就是说，集合中的元素存在互异性。集合set的这一特性使得在向集合添加元素的过程中，不需要检验集合中是否已经存在此元素。利用set数据结构可以存储一些集合性的数据，还可以实现交集、并集以及差集等运算操作。

5. zset（排序set）

zset（排序set）与set类似，是不包含相同字符串的集合。与set相比，zset增加了一个权重

① http://baike.so.com/doc/5063975.html.

参数，集合中的每个元素按照权重参数进行有序排列。有序集合zset的这一特性使得对集合进行添加、删除以及更新元素的操作速度很快，而且对有序集合的中间元素的访问速度也是非常快的。通常，有序集合被用来索引存储在Redis中的数据。

3.4 Column-oriented数据库

普通的关系型数据库都是以行为单位来存储数据的，擅长进行以行为单位的读入处理，而NoSQL的列数据库是以列为单位来存储数据的，因此擅长以列为单位读取数据。行数据库可以对少量行进行读取和更新，而列数据库可以对大量行少量列进行读取，并对于所有行和特定列进行同时更新。Column-oriented数据库具有高扩展性，即使增加数据也不会降低相应的处理速度。主要产品有Bigtable、Cassandra和HBase等。

3.4.1 Bigtable

Bigtable是Google设计的分布式数据存储系统，用来处理海量数据的一种非关系型的数据库。Bigtable是非关系型的数据库，是一个稀疏的、分布式的、持久化存储的多维度排序Map。Bigtable的设计目的是可靠地处理PB级别的数据，并且能够部署到上千台机器上。Bigtable已经实现的目标有：适用性广泛、可扩展、高性能和高可用性。目前Bigtable已经在超过60个Google产品和项目上得到了应用，其中就包括了Google Analytics、GoogleFinance、Orkut和GoogleEarth等。不同的产品和项目对Bigtable提出了不同的需求，有的需要高吞吐量的批处理，有的则需要及时响应，然后快速返回数据给最终用户。它们使用的Bigtable集群的配置也有很大的差异，有的集群只有几台服务器，有的则需要上千台服务器、存储几百TB的数据[①]。

1. Bigtable数据库的功能

Bigtable中有使用多数据库的实现策略，和数据库在很多方面类似。与此同时，内存数据库和并行数据库已经具备可扩展性和高性能，但Bigtable数据库却提供了一个和这些系统完全不一样的接口。Bigtable数据库不支持完整的关系数据模型，但它为客户提供了一个更简单的数据模型，客户通过这个模型可以动态地控制数据的分布和格式。数据的下标是行和列的名字，名字可以是任意字符串。Bigtable将存储的数据都视为字符串，但是它本身不去解释这些字符串，客户程序通常会在把各种结构化或者半结构化的数据串行化到这些字符串里。通过仔细选择数据的模式，客户可以控制数据的位置相关性。最后，通过Bigtable的模式参数来决定数据是存放在内存中还是存放在硬盘上。

2. Bigtable数据库的特点如下

- 适合海量PB级的数据。
- 支持动态伸缩，扩展容易。
- 适合读操作，不适用写操作。

① http://baike.baidu.com/view/3001038.htm?fr=aladdin.

- 分布式、并发数据处理，效率很高。
- 传统关系数据库不适用。
- 适用于廉价设备。

3. Bigtable数据库的应用

Bigtable数据库已在很多方面有着广泛的应用，比如，Bigtable数据库已为谷歌旗下的财经、地图、搜索，视频共享网站YouTube以及博客网站Blogger等提供了技术上的支持[1]。

3.4.2 Apache Cassandra

Apache Cassandra是一套开源分布式Key-Value存储系统。它最初由Facebook开发，用于储存特别大的数据。其主要特性表现在三个方面：分布式、基于column的结构化和高可扩展性。

Cassandra的主要特点是它不是一个数据库，而是由一堆数据库节点共同构成的一个分布式网络服务。对Cassandra的一个写操作，会被复制到其他节点上去；对Cassandra的读操作，会被路由到某个节点上去读取。对于一个Cassandra群集来说，扩展性能是比较简单的事情，只要在群集里添加节点就可以了。

Cassandra是一个混合型的非关系的数据库，类似于Google的BigTable，其主要功能比Dynomite（分布式的Key-Value存储系统）更丰富，但支持度却不如文档存储MongoDB（介于关系数据库和非关系数据库之间的开源产品，是非关系数据库中功能最丰富，最像关系数据库的。支持的数据结构非常松散，采用类似json的bjson格式，因此可以存储比较复杂的数据类型。）Cassandra是一个网络社交云计算方面理想的数据库。以Amazon专有的完全分布式的Dynamo为基础，结合了Google BigTable基于列族（Column Family）的数据模型，P2P去中心化的存储，很多方面都可以称之为Dynamo 2.0。

和其他数据库比较，Cassandra的突出特点如下。

- 模式灵活：使用Cassandra像文档存储，用户不必提前解决记录中的字段。用户可以在系统运行时随意的添加或移除字段。这是一个惊人的效率提升，特别是在大型部署上。
- 真正的可扩展性：Cassandra是纯粹意义上的水平扩展。为给集群添加更多容量，可以指向另一台电脑。用户不必重启任何进程，改变应用查询，或手动迁移任何数据。
- 多数据中心识别：用户可以调整节点布局来避免某一个数据中心起火，一个备用的数据中心将至少有每条记录的完全副本。
- 范围查询：如果用户不喜欢全部的键值查询，则可以设置键的范围来查询。
- 列表数据结构：在混合模式下可以将超级列添加到5维。对于每个用户的索引，这是非常方便的。
- 分布式写操作：可以在任何地方任何时间集中读或写任何数据，并且不会有任何单点失败。

3.4.3 HBase

HBase是基于Hadoop Distributed File System的一个开源的，采用列存储模型的分布式数据

---

[1] http://baike.baidu.com/view/3001038.htm?fr=aladdin.

库产品。它是一个可以利用Hadoop HDFS作为其文件存储系统，提供高可靠性、高性能、列存储、可伸缩以及实时读写的数据库系统。它利用Hadoop MapReduce来处理HBase中的海量数据，利用Zookeeper作为协同服务。HBase在Hadoop Ecosystem中的位置如图3.11所示。

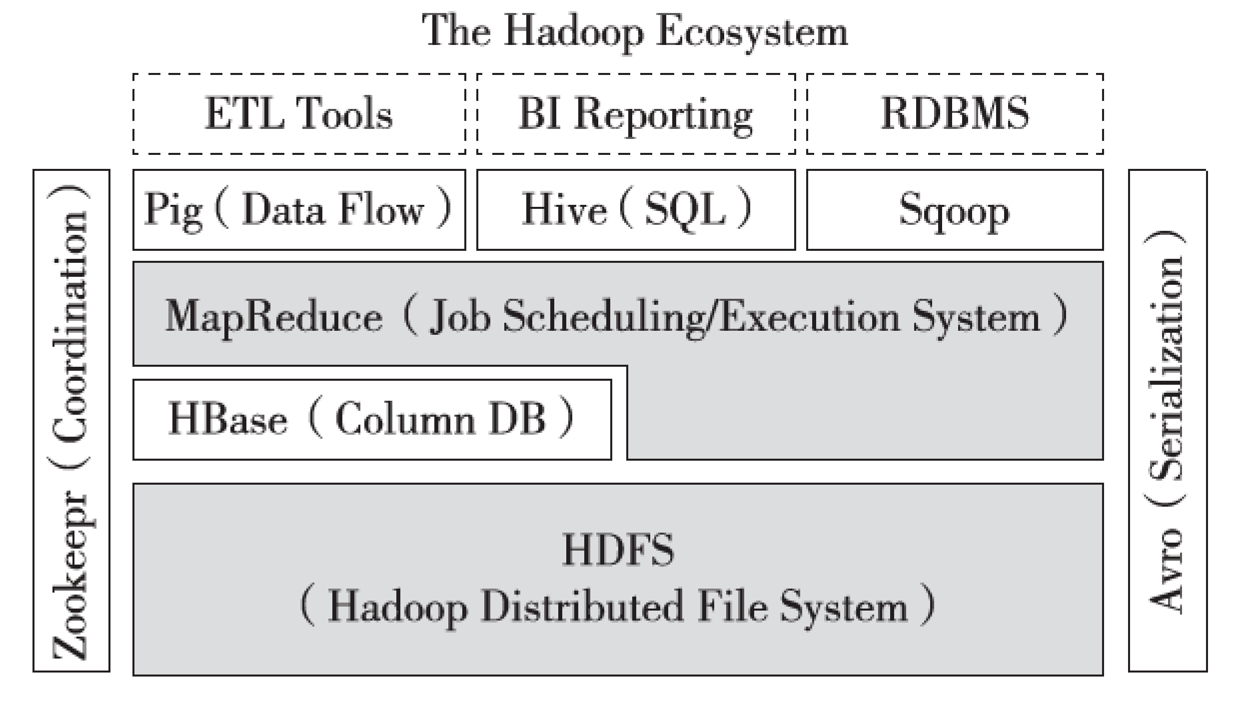


图3.11　HBase在Hadoop Ecosystem中的位置

HBase中表的特点主要表现在以下五个方面[①]。

- 大：单个表可以有上亿行，上百万列。
- 面向列：面向列（族）的存储和权限控制，列（族）独立检索。
- 稀疏：对于为空（null）的列，不占用存储空间。因此，表也可以设计得非常稀疏。
- 每个cell中的数据可以有多个版本，默认情况下版本号自动分配，是单元格插入时的时间戳。
- HBase中的数据都是字符串，没有类型。

HBase以表的形式存储数据。表由行和列组成。列划分为若干个列族（row family），因此HBase的逻辑视图如表3.3所示。

表3.3　HBase的逻辑视图

RowKey	column–family1		column–family2			column–family3
	column1	column2	column1	column2	column3	column1
key1	t1:abc t2:gdxdf		t4:dfads t3:hello t2:world			
key2	t3:abc t1:gdxdf		t4:dfads t3:hello		t2:dfdsfa t3:dfdf	
key3		t2:dfadfasd t1:dfdasddsf				t2:dfxxdfasd t1:taobao.com

1. HBase中数据表的物理存储方式

在这里，我们将介绍HBase中数据表的物理存储方式，具体如下所示。

- 表中的所有行都按照Row Key的字典序排列。
- 表在行方向上分割为多个Region。
- Region按大小分割的，每个表一开始只有一个Region，但随着数据的不断插入表，Region不断会增大。当增大到一个阀值时， Region就会等分成两个新的Region。当

① http://blog.csdn.net/anghlq/article/details/6538229

表中的行不断增多，就会有越来越多的Region与之相对应，如图3.12和图3.13所示。

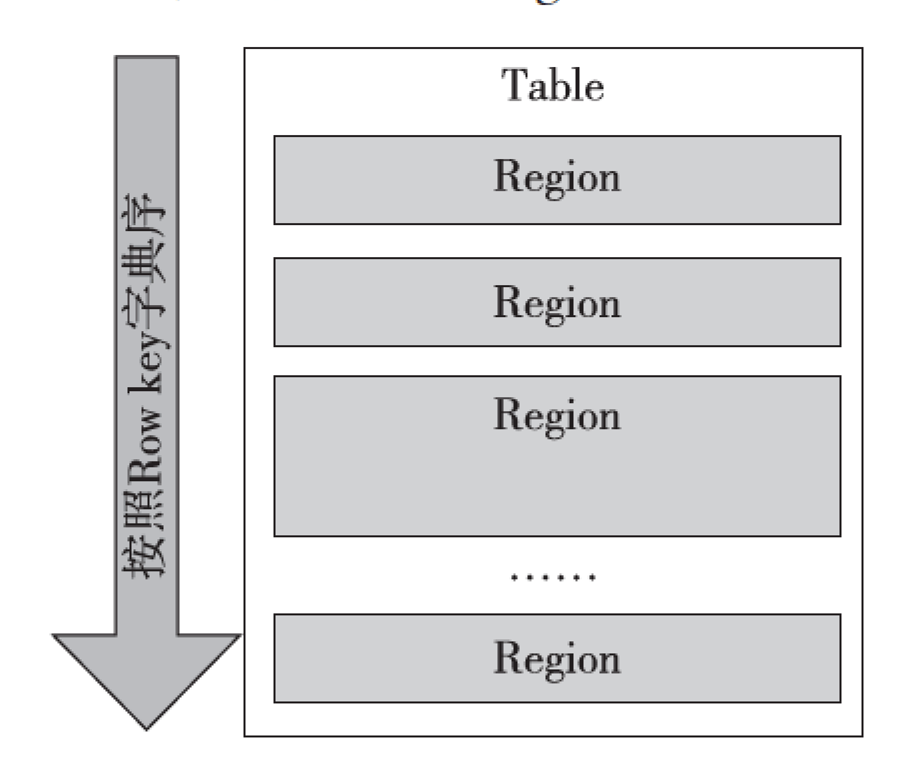


图3.12　HBase数据表存储方式

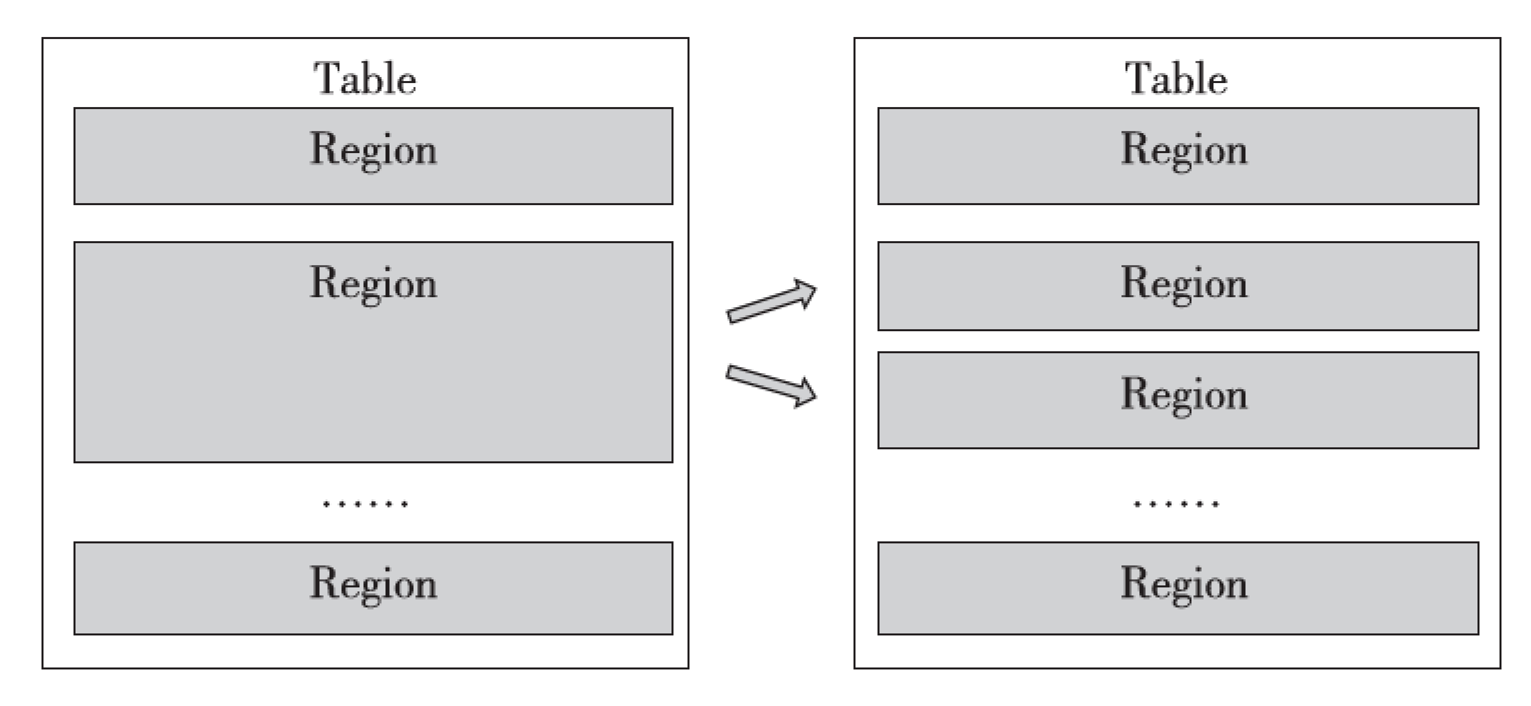


图3.13　分到两个新的Region

- Region是HBase中分布式存储和负载均衡的最小单元。最小单元是指不同的Region可以分布在不同的Region Server上。但是一个Region是不会拆分到region server，如图3.14所示。

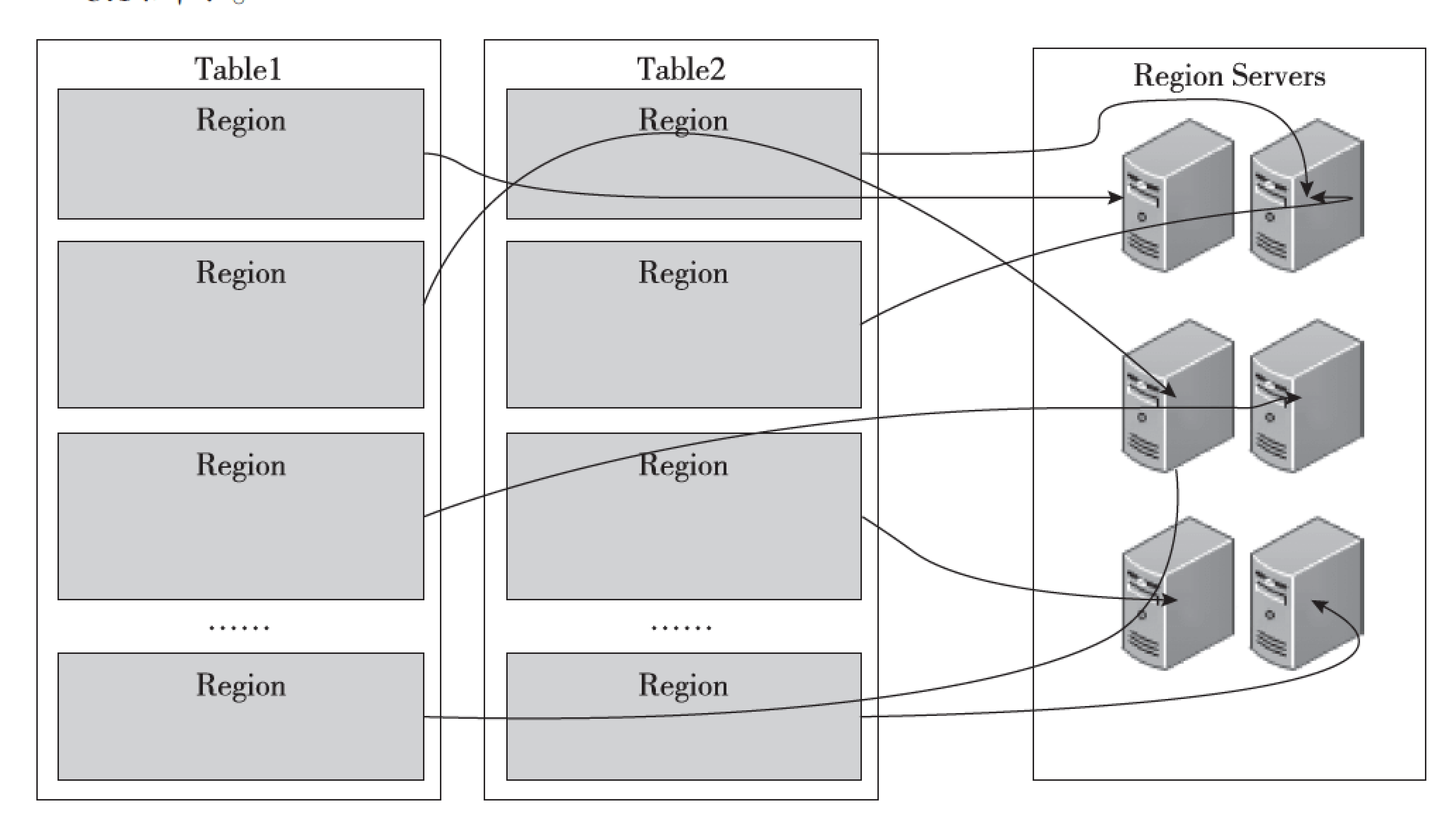


图3.14　Region Server与表

● Region虽然是分布式存储的最小单元，但它并不是存储的最小单元。事实上，Region由一个或多个Store组成，每个Store保存一个columns family，而每个Store又是由一个memStore和0到多个StoreFile组成，如图3.15所示。

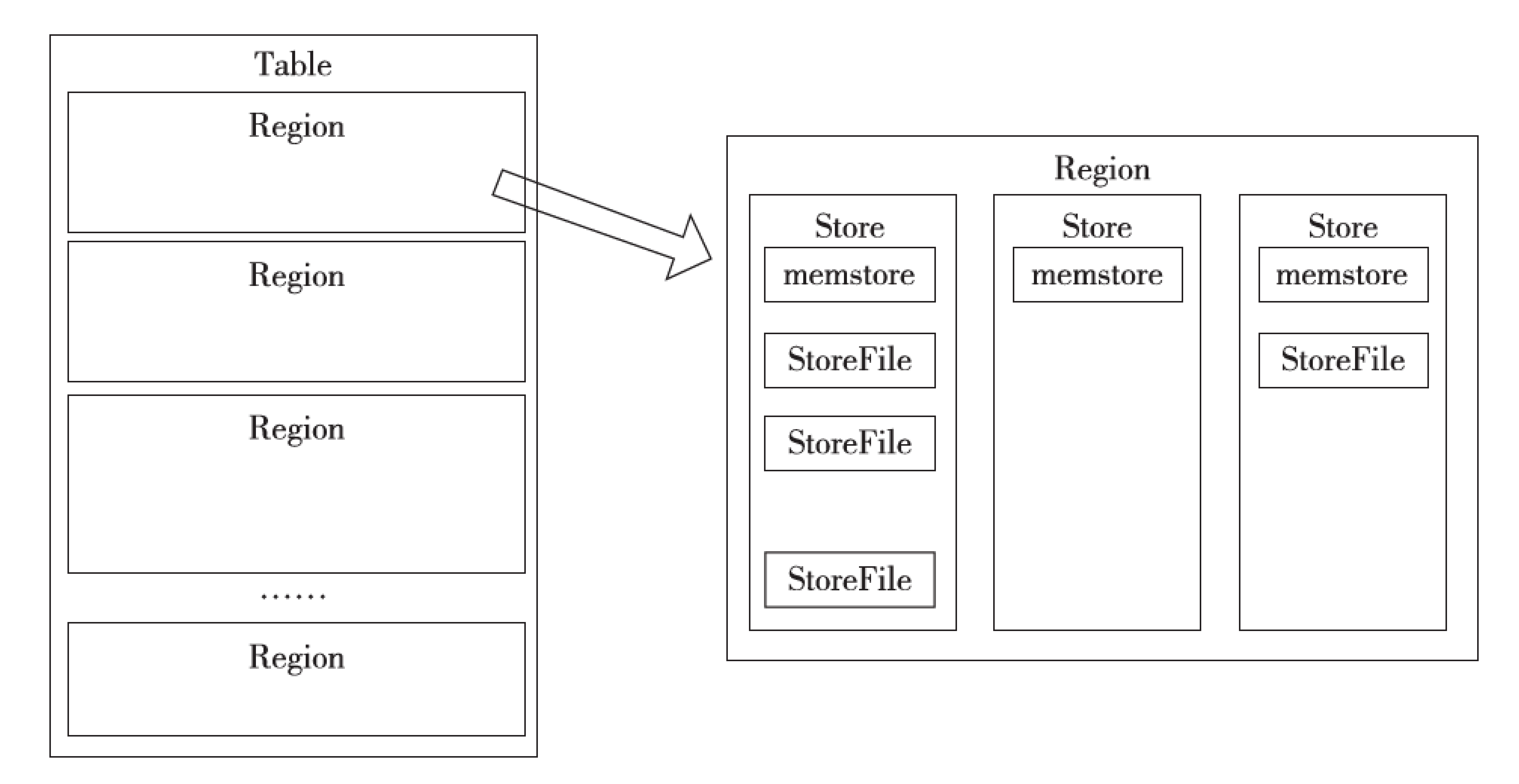


图3.15　Region的存储组成单元

HBase的系统架构可以参照图3.16和图3.17[①]所示。

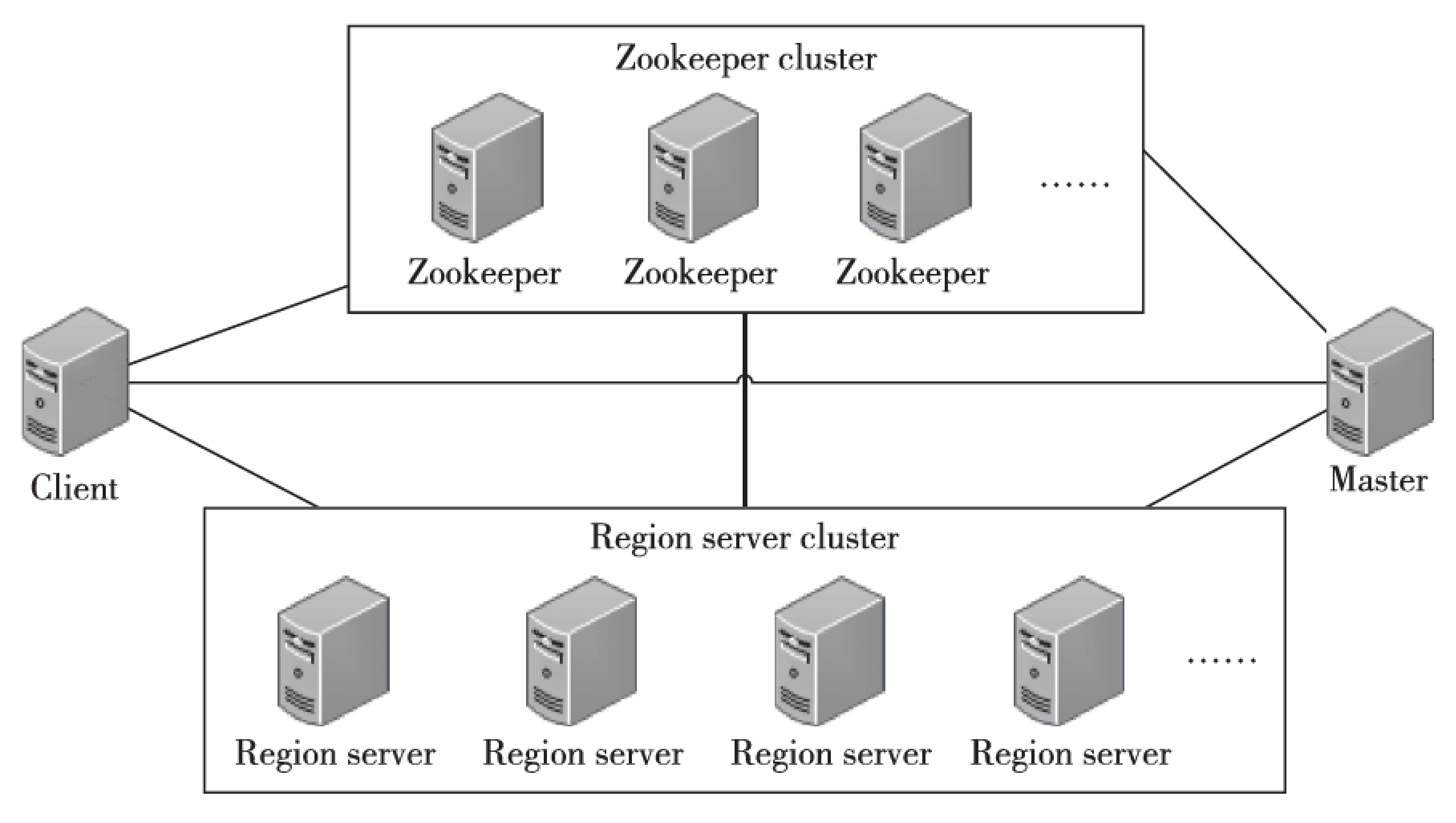


图3.16　HBase系统架构

（1）Client

Client包含访问HBase的接口，同时也维护着一些cache，来加快对HBase的访问，如Region的位置信息。

（2）Zookeeper

● 保证在任何时候，集群中有且仅有一个master。
● 存储所有的Region寻址入口。

① http://blog.csdn.net/anghlq/article/details/6538229

● 对Region Server的状态实时监控，并及时将Region Server的上线和下线信息通知给Master。
● 对HBase的schema进行存储，包括有哪些表以及每个表有哪些column family。

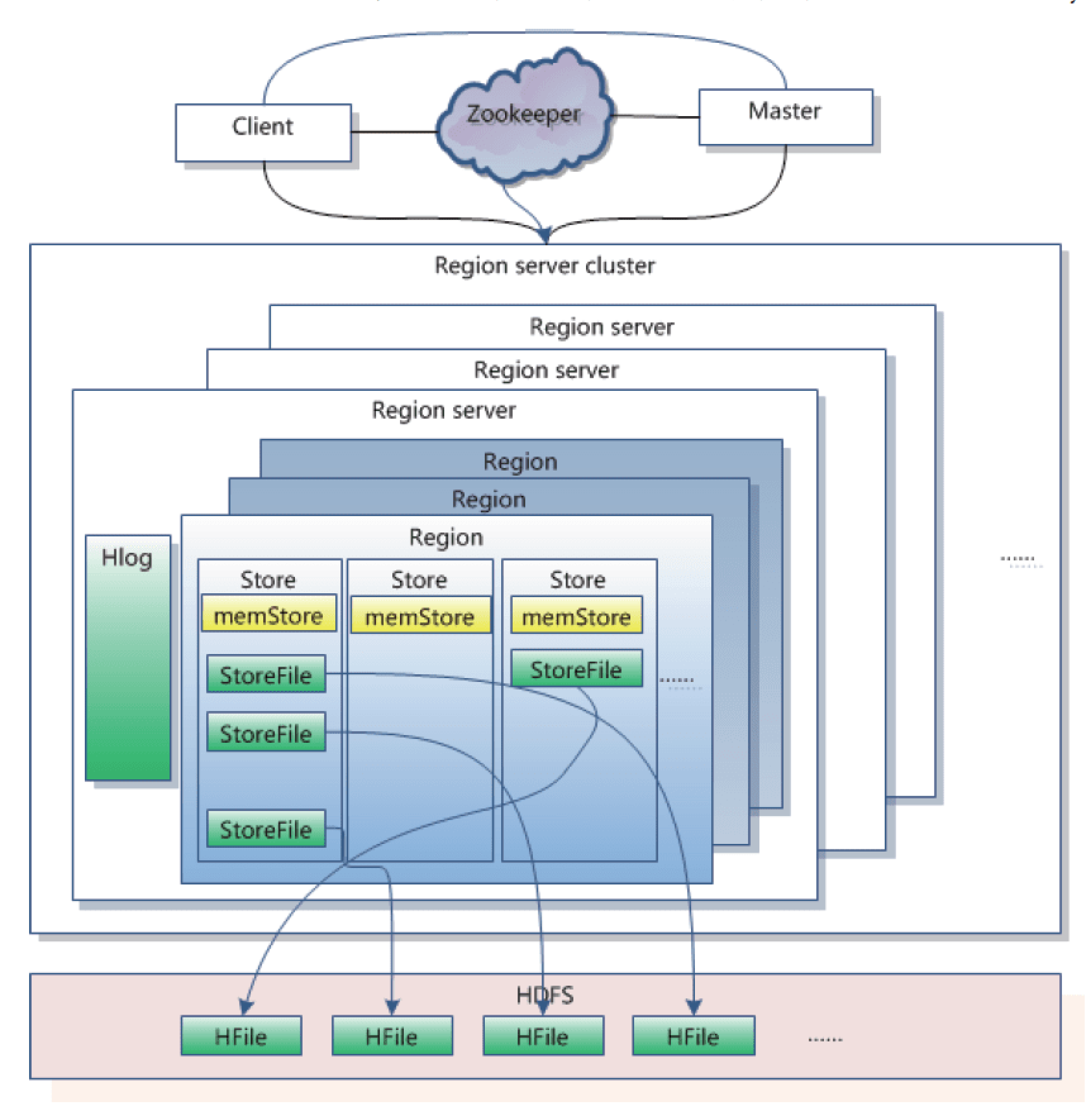


图3.17　HBase系统架构

（3）Master

HBase集群中通常有一个master节点，其功能作用如下。

● 为Region Server分配Region。
● 负责Region Server的负载，使其均衡存在。
● 若发现失效的Region Server，重新分配Region。
● 对GFS上的垃圾文件进行回收。
● 对schema的更新请求进行相应的处理。

（4）Region Server

RegionServer是HBase集群运行在各个工作节点之上的服务，它是整个HBase系统的关键所在，其功能作用如下。

● 维护Master分配给它的Region，并处理对这些Region的I/O请求。
● 负责切分在运行过程中变得过大的Region。

可以看到，Client在访问HBase上的数据时，并不需要Master的参与（寻址访问Zookeeper

和Region Server，数据读写访问Region Server），Master的作用仅仅是维护表和Region的元数据信息，并且它的负载往往很低。

2. HBase的关键算法及流程

（1）Region的定位

系统要怎样才能找到某个Row Key或者Row Key range所在的Region。HBase提供了一个三层类似B+树的结构来保存Region的位置。

第一层持有Root Region的位置，负责储存Zookeeper里的文件。

第二层Root Region是.META.表的第一个Region，用户可以通过Root Region来访问.META.表中的数据。它将.META.z表中其他Region的位置保存了下来。

第三层是.META.，如图3.18所示，它将HBase中所有数据表的Region位置信息保存了下来，是一个特殊的表。

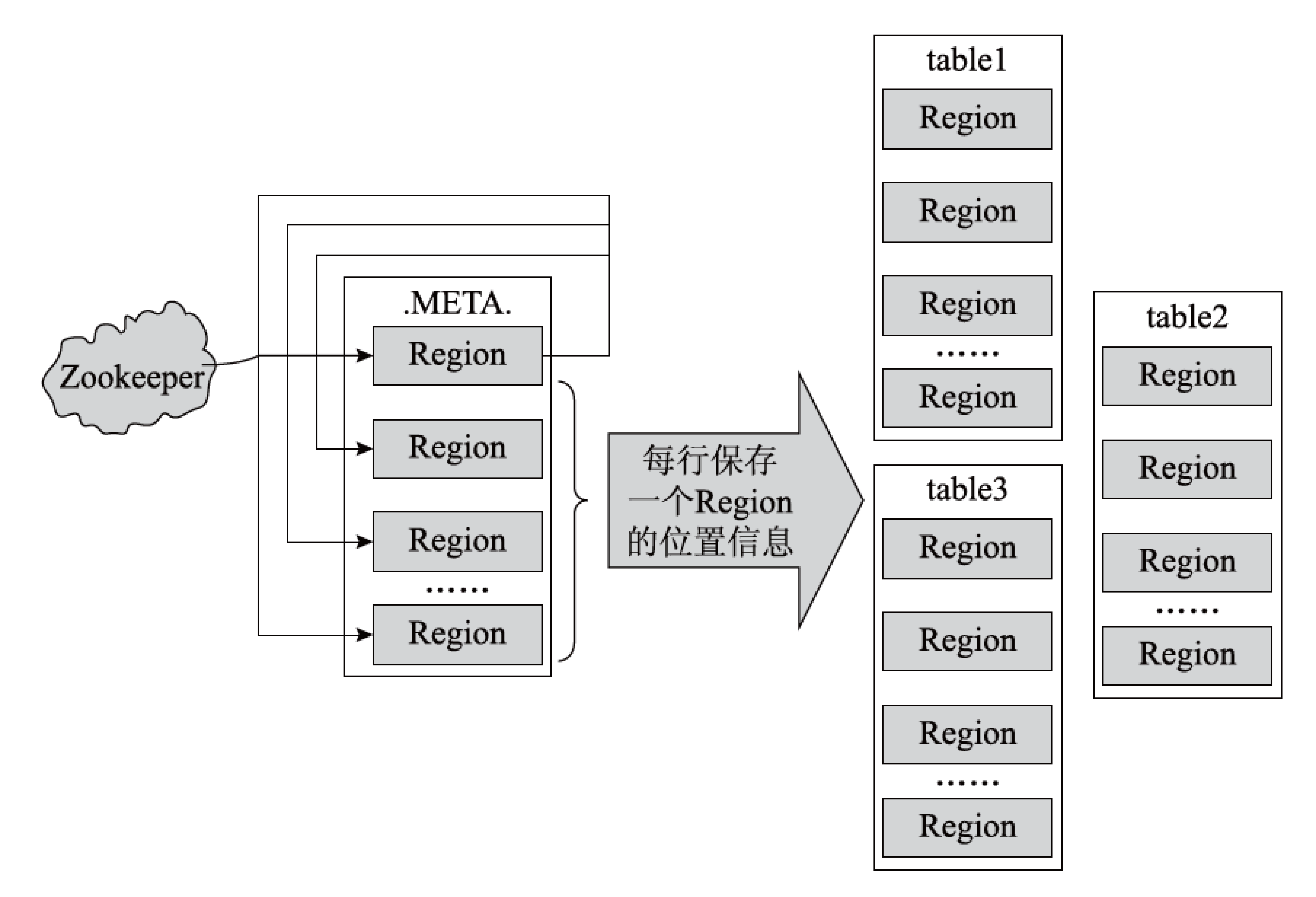


图3.18 HBase的关键算法及流程

图3.18 HBase的关键算法及流程说明如下。

a. 在.META.表中，每一行都保存了一个Region的位置信息，可由表名+表的最后一样编码形成Row Key。

b. 不会发生Root Region被split的情况，可以保证三次跳转后定位到任意Region。

c. client的作用是把查询过的位置信息存入缓存，并且缓存不会主动失效。假如client上的缓存全部失效，需要进行6次网络的来回，才能重新定位到正确的Region（其中三次用来发现缓存失效，而另外三次则用来获取位置信息）[①]。

d. 在内存中保存了.META.表的全部Region，目的是为了提高访问的速度。

（2）读写过程

前面提到，HBase对表进行更新主要是通过MemStore和StoreFile存储来进行的。在更新

① http://blog.csdn.net/anghlq/article/details/6538229

时，第一步先在内存（MemStore）和Log（WAL log）中写入数据。其中，数据在MemStore中是有排序的，当MemStore中的数据到达某一阈值时，一个新的MemStore就会被创建出来，原来的MemStore会添加到flush队列中，通过单独的线程flush到磁盘上，形成一个StoreFile。同时，系统会在Zookeeper中记录一个redo point，表示此刻之前的变更已经持久化。如若系统出现意外，内存中的数据可能会丢失，这时可以通过使用Log（WAL log）将checkpoint之后的数据恢复。而StoreFile是只读的，形成后就不能修改，因此更新HBase就是在不断地追加操作。当一个Store中的StoreFile达到某一阈值时，就会将同一个key修改合并到一起，然后形成一个大的StoreFile，当StoreFile达到某一阈值时，会等分成两个StoreFile。由于更新表的这个过程是不断追加的，因此对读请求进行处理时，Store中全部的MemStore和StoreFile都会被访问，然后按照Row Key合并MemStore和StoreFile。在合并的过程中，由于StoreFile和MemStore都是经过排序的，并且StoreFile带有内存索引，所以所需时间较少。

写请求的处理过程如图3.19所示。

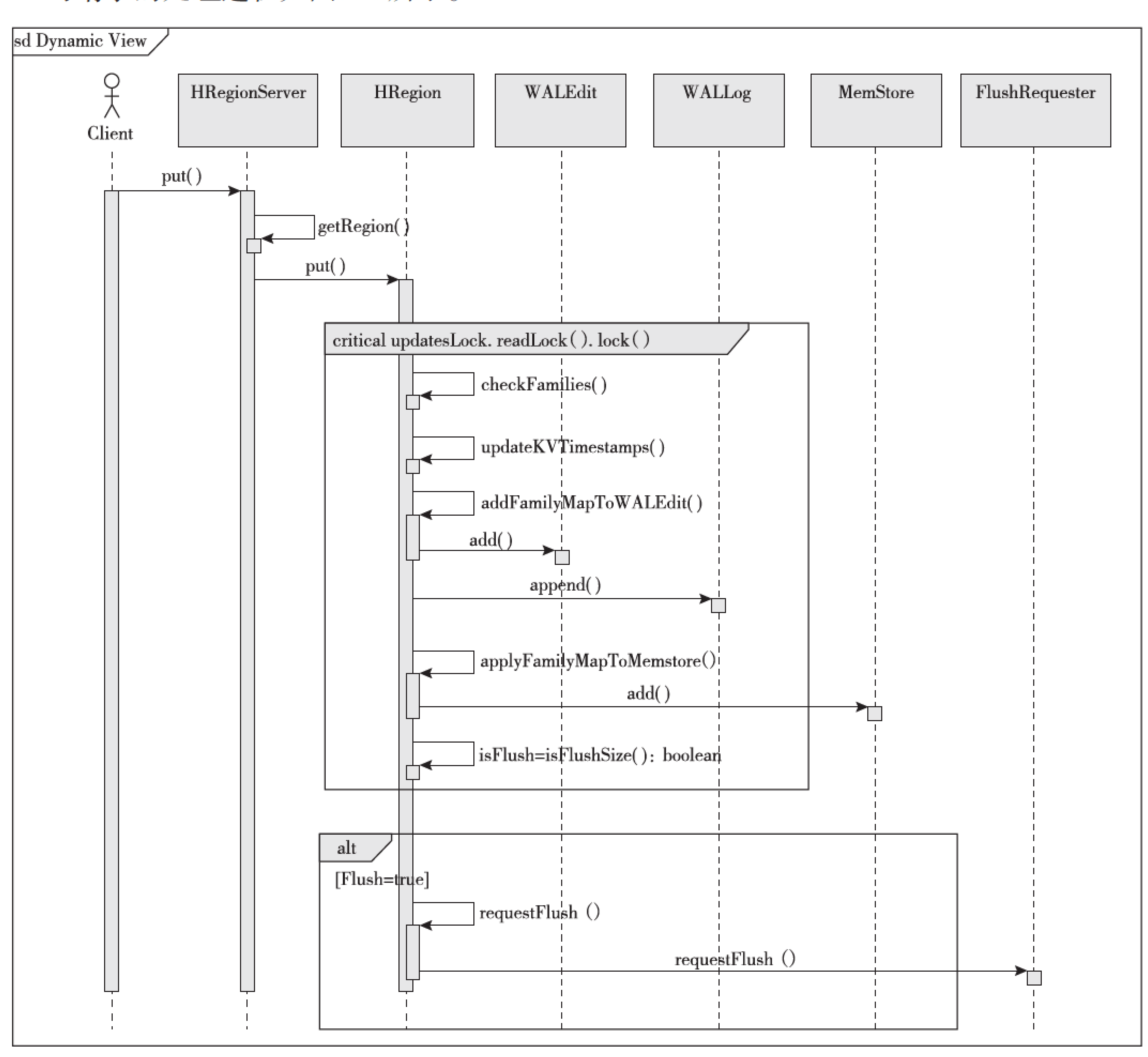


图3.19 写请求处理过程

a. Client首先向Region Server提交写请求。

b. Region Server找到目标Region。

c. Region检查数据是否与schema相一致。

d. 如果客户端没有规定某个特定的版本，则可以获取当前系统时间作为数据版本。

e. 将更新写入WAL log中。

f. 将更新写入Memstore中。

g. 判断Memstore是否需要flush为Store文件。

（3）Region的分配

一个Region Server在任何时刻只能分配到一个Region。Master则记录了当前有哪些Region Server可用，哪些Region还没有分配，以及当前哪些Region分配给了哪些Region Server。当有Region还没有分配，同时存在一个Region Server上有可用空间，Master就会给这个Region Server发送一个装载请求，Region Server收到请求后，就可以开始对该Region提供服务，把还没有被分配的Region分配给这个有可用空间的Region Server。

（4）Region Server上线

为了跟踪Region Server的状态，Master借助于Zookeeper。启动某个Region Server时，在Zookeeper的Server目录下Region Server首先会建立代表它自己的文件，同时获得该文件的独占锁。由于Master订阅了Server目录上的变更消息，所以当出现新增或删除的操作在Server目录下的文件时，Master会通过Zookeeper立刻知道。因此，只要Region Server上线，Master就能立刻知道消息。

（5）Region Server下线

当Region Server上线时，Zookeeper会获得代表这台Server的文件上的独占锁。相应地，当Region Server下线时，Region Server和Zookeeper的会话就会断开，Zookeeper会自动释放独占锁。Master会不断地查询Server目录下文件的锁状态，一旦Master发现某个Region Server的独占锁丢失了（或者Master和Region Server通信连续几次都不成功），则Master就会尝试着去获取代表这个Region Server的读写锁。一旦获取成功，即可确定以下两种情况发生了：Region Server与Zookeeper之间的网络断开了，或者Region Server挂了。无论哪种情况发生，Region Server都将无法继续为它的Region提供服务，此时Master会将Server目录下代表这台Region Server的文件删除，并重新分配这台Region Server的Region。

如果Region Server丢失了它的锁是因为网络暂时出了问题，那么Region Server重新连接到Zookeeper之后，只要代表它的文件没有丢失，它就会不断地尝试获取这个文件上的锁。一旦获取到了，就可以继续为其提供服务。

（6）master上线

启动Master需进行以下步骤。

a. 首先要阻止其他Master成为Master，因此可从Zookeeper上获取惟一一个代码Master的锁。

b. 对Zookeeper上的Server目录进行扫描，以便得到当前可用的Region Server列表。

c. 与步骤b中的每个Region Server进行通信，从而获得当前已分配的Region以及Region Server的对应关系。

d. 对.META.Region集合进行扫描，通过计算得到当前还没有分配的Region，并将它们放入待分配的Region列表中。

（7）Master下线

Master并不参与表的数据I/O的过程，只会维护Region和表的元数据。Master下线会导致所有元数据的修改被冻结，而读写表的数据还是可以正常进行的，因此在短时间内Master下线对整个HBase集群并没有显著的影响。从上线过程可以看到，Master保存的信息都可以从系统其他地方收集到或计算出来，因此，在一般的HBase集群中，总会有一个Master在提供服务，同时也存在着另一个或一个以上的“master”在等待时机抢占它的位置。

3.5　图存数据库

图存数据库主要是将数据以图的方式存储。它主要适用于关系较强的数据中，但适用范围很小，因为很少有操作涉及到整个图。主要产品有Neo4j、GraphDB和OrientDB等，本书主要介绍Neo4j。

3.5.1　Neo4j

Neo4j是一个用Java实现的、完全兼容ACID的图形数据库。数据以一种针对图形网络进行过优化的格式保存在磁盘上。Neo4j的内核是一种极快的图形引擎，具有数据库产品期望的所有特性，如恢复、两阶段提交、符合XA等。

Neo4j是图形数据库中最重要的一种。它使用数据结构中图（graph）的概念来进行建模。Neo4j中两个最基本的概念是节点和边。节点表示实体，边则表示实体之间的关系。节点和边都可以有自己的属性。不同实体通过各种不同的关系关联起来，形成复杂的对象图。Neo4j同时提供了在对象图上进行查找和遍历的功能。Neo4j使用“图”这种最通用的数据结构来对数据进行建模，使得Neo4j的数据模型在表达能力上非常强。链表、树和散列表等数据结构都可以抽象成图来表示。Neo4j同时具有一般数据库的基本特性，包括事务支持、高可用性和高性能等。Neo4j已经在很多生产环境中得到了应用。流行的云应用开发平台Heroku也提供了Neo4j作为可选功能。

对于很多应用来说，领域对象模型本身就是一个图结构。对于这样的应用，使用Neo4j这样的图形数据库进行存储是最适合的，因为在进行模型转换时代价最小。以基于社交网络的应用为例，用户作为应用中的实体，通过不同的关系关联在一起，如亲人关系、朋友关系和同事关系等。不同的关系有不同的属性。比如同事关系所包含的属性包括所在公司的名称、开始的时间和结束的时间等。对于这样的应用，使用Neo4j来进行数据存储，不仅实现起来简单，后期的维护成本也比较低。

1. Neo4j的基本使用

Neo4j的基本使用包括节点和关系的使用，Neo4j能够对节点进行索引，同时还支持非常复杂的图的遍历操作。

（1）节点和关系

Neo4j中最基本的概念是节点（node）和关系（relationship）。节点表示实体，由org.neo4j.graphdb.Node接口来表示。在两个节点之间，可以有不同的关系。关系由org.neo4j.graphdb.Relationship接口来表示。每个关系由起始节点、终止节点和类型等三个要素组成。起始节点和终止节点的存在，说明了关系是有方向，类似于有向图中的边。不过在某些情况下，关系的方向可能并没有意义，会在处理时被忽略。所有的关系都是有类型的，用来区分节点之间意义不同的关系。在创建关系时，需要指定其类型。关系的类型由org.neo4j.graphdb.RelationshipType接口来表示。节点和关系都可以有自己的属性。每个属性是一个简单的名值对。属性的名称是String类型的，而属性的值则只能是基本类型、String类型以及基本类型和String类型的数组。一个节点或关系可以包含任意多个属性。对属性进行操作的方法声明在接口org.neo4j.graphdb.PropertyContainer中。Node和Relationship接口都继承自PropertyContainer接口。PropertyContainer接口中常用的方法包括获取和设置属性值的getProperty和setProperty。下面通过具体的示例来说明节点和关系的使用方法。

此示例是一个简单的歌曲信息管理程序，用来记录歌手、歌曲和专辑等相关信息。在这个程序中，实体包括歌手、歌曲和专辑，关系则包括歌手与专辑之间的发布关系以及专辑与歌曲之间的包含关系。

例一：节点和关系的使用示例

```
private static enum RelationshipTypes implements RelationshipType {
    PUBLISH, CONTAIN
 }

public void useNodeAndRelationship() {
    GraphDatabaseService db = new EmbeddedGraphDatabase("music");
    Transaction tx = db.beginTx();
    try {
        Node node1 = db.createNode();
        node1.setProperty("name", "歌手 1");
        Node node2 = db.createNode();
        node2.setProperty("name", "专辑 1");
        node1.createRelationshipTo(node2, RelationshipTypes.PUBLISH);
        Node node3 = db.createNode();
        node3.setProperty("name", "歌曲 1");
        node2.createRelationshipTo(node3, RelationshipTypes.CONTAIN);
        tx.success();
    } finally {
        tx.finish();
    }
 }
```

在例一中首先定义了两种关系类型。定义关系类型的一般做法是创建一个实现了RelationshipType接口的枚举类型。RelationshipTypes中的PUBLISH和CONTAIN分别表示发布和包含关系。在Java程序中可以通过嵌入的方式来启动Neo4j数据库，只需要创建org.neo4j.kernel.EmbeddedGraphDatabase类的对象，并指定数据库文件的存储目录即可。在使用Neo4j数据库时，进行修改的操作一般需要包含在一个事务中进行处理。通过GraphDatabaseService接口的createNode方法可以创建新的节点。Node接口的createRelationshipTo方法可以在当前节点和另外一个节点之间创建关系。

另外一个与节点和关系相关的概念是路径。路径有一个起始节点，接着的是若干个成对的关系和节点对象。路径是在对象图上进行查询或遍历的结果。Neo4j中使用org.neo4j.graphdb.Path接口来表示路径。Path接口提供了对其中包含的节点和关系进行处理的一些操作，包括使用startNode和endNode方法来获取起始和结束节点，以及使用nodes和relationships方法来获取遍历所有节点和关系的Iterable接口的实现。关于图上的查询和遍历，下面还会有介绍。

（2）使用索引

当Neo4j数据库中包含的节点比较多时，要快速查找满足条件的节点会比较困难。Neo4j提供了对节点进行索引的能力，可以根据索引值快速地找到相应的节点。下面通过具体的示例来说明索引的基本用法。

例二：使用的索引示例

```
public void useIndex() {
    GraphDatabaseService db = new EmbeddedGraphDatabase("music");
    Index<Node> index = db.index().forNodes("nodes");
    Transaction tx = db.beginTx();
    try {
        Node node1 = db.createNode();
        String name = "歌手 1";
        node1.setProperty("name", name);
        index.add(node1, "name", name);
        node1.setProperty("gender", "男");
        tx.success();
    } finally {
        tx.finish();
    }
    Object result = index.get("name", "歌手 1").getSingle()
 .getProperty("gender");
    System.out.println(result); // 输出为“男”
 }
```

例二通过GraphDatabaseService接口的index方法可以得到管理索引的org.neo4j.graphdb.index.IndexManager接口的实现对象。Neo4j支持对节点和关系进行索引。通过IndexManager

接口的forNodes和forRelationships方法可以分别得到节点和关系上的索引。索引通过org.neo4j.graphdb.index.Index接口来表示，其中的add方法用来把节点或关系添加到索引中，get方法用来根据给定值在索引中进行查找。

（3）图的遍历

在图上进行的最实用的操作是图的遍历。通过遍历操作，可以获取与图中节点之间的关系相关的信息。Neo4j支持非常复杂的图的遍历操作。在进行遍历之前，需要对遍历的方式进行描述。遍历的方式的描述信息由下列几个要素组成。

- 遍历的路径：通常用关系的类型和方向来表示。
- 遍历的顺序：常见的遍历顺序有深度优先和广度优先两种。
- 遍历的惟一性：可以指定在整个遍历中是否允许经过重复的节点、关系或路径。
- 遍历过程的决策器：用来在遍历过程中判断是否继续进行遍历，以及选择遍历过程的返回结果。
- 起始节点：遍历过程的起点。

Neo4j中遍历方式的描述信息由org.neo4j.graphdb.traversal.TraversalDescription接口来表示。通过TraversalDescription接口的方法可以描述上面介绍的遍历过程的不同要素。类org.neo4j.kernel.Traversal提供了一系列的工厂方法用来创建不同的TraversalDescription接口的实现。例三中给出了进行遍历的示例。

例三：遍历操作的示例

```
TraversalDescription td = Traversal.description()
    .relationships(RelationshipTypes.PUBLISH)
    .relationships(RelationshipTypes.CONTAIN)
    .depthFirst()
    .evaluator(Evaluators.pruneWhereLastRelationshipTypeIs(RelationshipTypes.
CONTAIN));
Node node = index.get("name", "歌手 1").getSingle();
Traverser traverser = td.traverse(node);
for (Path path : traverser) {
    System.out.println(path.endNode().getProperty("name"));
}
```

在例三中，首先通过Traversal类的description方法创建了一个默认的遍历描述对象。通过TraversalDescription接口的relationships方法可以设置遍历时允许经过的关系的类型，而depthFirst方法用来设置使用深度优先的遍历方式。比较复杂的是表示遍历过程的决策器的evaluator方法。该方法的参数是org.neo4j.graphdb.traversal.Evaluator接口的实现对象。Evalulator接口只有一个方法evaluate。evaluate方法的参数是Path接口的实现对象，表示当前的遍历路径，而evaluate方法的返回值是枚举类型org.neo4j.graphdb.traversal.Evaluation，表示不同的处理策略。处理策略由两个方面组成：第一个方面为是否包含当前节点，第二个方面为是否继续进行遍历。Evalulator接口的实现者需要根据遍历时的当前路径，做出相应的决策，返回适当

的Evaluation类型的值。类org.neo4j.graphdb.traversal.Evaluators提供了一些实用的方法来创建常用的Evalulator接口的实现对象。例三中使用了Evaluators类的pruneWhereLastRelationshipTypeIs方法。该方法返回的Evalulator接口的实现对象会根据遍历路径的最后一个关系的类型来进行判断，如果关系类型满足给定的条件，则不再继续进行遍历。

例三中的遍历操作的作用是查找一个歌手所发行的所有歌曲。遍历过程从表示歌手的节点开始，沿着RelationshipTypes.PUBLISH和RelationshipTypes.CONTAIN这两种类型的关系，按照深度优先的方式进行遍历。如果当前遍历路径的最后一个关系是RelationshipTypes.CONTAIN类型，则说明路径的最后一个节点包含的是歌曲信息，可以终止当前的遍历过程。通过TraversalDescription接口的traverse方法可以从给定的节点开始遍历。遍历的结果由org.neo4j.graphdb.traversal.Traverser接口来表示，可以从该接口中得到包含在结果中的所有路径。结果中的路径的终止节点就是表示歌曲的实体。

3.6 文档数据库

文档数据库是一种用来管理文档的数据库，它与传统数据库的本质区别在于，其信息处理基本单位是文档，可长、可短、甚至可以无结构。它与关系数据库的主要区别在于，文档数据库允许建立不同类型的非结构化或者任意格式的字段，并且不提供完整性的支持。但是它与关系型数据库并不是相互排斥的，它们之间可以相互补充、扩展。文档数据库的两个典型代表是CouchDB和MongoDB。

3.6.1 CouchDB

CouchDB[①]（Couch是cluster of unreliable commodity hardware的首字母缩写）是前IBM公司Lotus Notes开发者Damien Katz创建于2005年的一个项目。Damien Katz将其定义为“面向大规模可扩展对象数据库的存储系统”。他对该数据库的目标是让其成为互联网的数据库，因此其从底层的设计就支持部署Web应用程序。Damien Katz自己在近两年的时间内创建该项目的同时，也将其作为开源项目在GNU General Public License上发布。在2008年2月，该项目成为Apache Incubartor项目，同时其许可协议变更为Apache License。几个月后，CouchDB升级为顶级项目，这促使其第一个稳定版在2010年7月发布。在2012年初期，Damien Katz离开了该项目而专注于CouchBase Server。不过该项目还在继续进行中，在2012年4月发布了1.2版本，2013年4月发布了1.3版本。

CouchDB的主要功能是将数据存储为“文档”，内容为用JSON表示的有一个或者多个字段/值的对。这里JSON的数据类型和结构可以参见表3.4。CouchDB中的每一个文档有一个惟一的ID但是没有必须的文档schema。

此外，CouchDB提供了ACID语义，意味着CouchDB能够处理大量的并发读写而不会产生冲突。同时，在CouchDB中，每一个视图都是由作为map/reduce操作中的Map部分的JavaScript函数构成。该函数接受一个文档并且将其转换为一个单独的值来返回。CouchDB能够对视图

① http://baike.so.com/doc/5903671.html.

进行索引，同时在文档新增、修改、删除的时候对这些索引进行更新。CouchDB还支持复制的分布式架构。CouchDB的设计基于支持双向的复制（同步）和离线操作。这意味着多个复制能够对同一数据有其自己的副本，可以进行修改，之后，同步这些变更。

表3.4　JSON各类型值示例

类 型	示 例	示 例	示 例
数值	"count"3	"acore" : –15	"height": 13.21
字符串	"database": "CouchDB"	"context": "can I help you?"	
布尔	"flag": "false"	"flag": "true"	
数组	"运动"：{"篮球"，"足球"，"游泳"}	[{"运动"：["篮球", "足球", 游泳]},{"height":1.80},{"context": "can I help you?"}	
对象	{"姓名": "李刚", "性别": "男", "生日": "1990–01–12"}		
空值	"Email": null		

CouchDB可以保证最终一致性，使其能够同时提供可用性和分割容忍。此外，CoucbDB能够同步复制到可能会离线的终端设备（比如智能手机）中，当设备再次在线时再处理数据同步。

1. CouchDB的特点①

- CouchDB与lucene的index结构相类似，是面向文档的数据库，对半结构化的数据进行存储，特别适合存储文档，因此很适合电话本、CMS、地址本等的应用。在这些应用中，相较于关系数据库，文档数据库要更加方便，性能也更好。
- CouchDB是一个分布式的数据库，需要依靠Erlang无与伦比的并发特性，在n台物理节点上对存储系统进行分布，并且很好地对节点之间的数据读写一致性进行同步和协调。在应用基于Web的大规模应用文档时，分布式的数据库可以大量地改动应用代码层，使它不必像传统的关系数据库那样还要分库拆表。
- CouchDB支持REST API，操作CouchDB数据库时可以让用户使用JavaScript来进行，也可以在编写查询语句时使用JavaScript。借助于此，用AJAX技术结合CouchDB开发出来的CMS系统更加简单方便。

2. 工作原理

CouchDB的构建基础是在强大的B–树储存引擎上。该引擎提供一种能够在对数均摊时间内执行搜索、插入和删除操作的机制，并负责排序CouchDB中的数据。B–树储存引擎被用于所有的内部文档、数据和视图。

由于CouchDB数据库的结构独立于模式，所以它在创建文档之间的任意关系，以及提供聚合和报告特性时是依赖于视图的。Map/Reduce是一种使用分布式计算来处理和生成大型数据集的模型，对这些视图的结果是需要使用Map/Reduce来计算。Map/Reduce模型主要由Google引入，分为两个步骤。在Map步骤中，主节点对文档进行接收，然后将问题划分为多个子问题。然后将这些子问题发布给工作节点，由工作节点处理后再将结果返回给主节点。而在Reduce步骤中，主节点对来自工作节点的结果进行接收并把它们合并，从而获得能够解决

① http://baike.so.com/doc/5903671.html.

最初问题的总体结果和答案。

CouchDB中的Map/Reduce特性是生成键/值对，它们被CouchDB插入到B–树引擎中并且可以根据它们的键进行排序。用户可以通过键进行高效的查找，并提高B–树中的操作的性能。此外，这还意味着对数据进行分区时可以在多个节点上，而不需要单独查询每个节点。

传统的关系数据库管理系统对并发性的管理有时是通过锁来进行，从而防止某个客户机正在更新的数据被其他客户机访问。

这就防止多个客户机同时更改相同的数据，但对于多个客户机同时使用一个系统的情况，数据库在确定哪个客户机应该接收锁并维护锁队列的次序时会遇到困难，这很常见。在CouchDB中没有锁机制，它使用的是多版本并发性控制（Multiversion concurrency control，MVCC），它向每个客户机提供数据库的最新版本的快照。这意味着在提交事务之前，其他用户不能看到更改。许多现代数据库开始从锁机制前移到MVCC，包括Oracle（V7之后）和Microsoft® SQL Server 2005及更新版本。

3. 最佳应用场景

适合数据变化相对较少，进行数据统计，执行预定义查询的应用程序。对需要提供数据版本支持的应用程序较适用。

3.6.2 MongoDB

MongoDB[①]是一个介于关系与非关系数据库之间的产品，在非关系数据库之中它的功能最为丰富，与关系数据库最为接近。它支持的是一种类似于JSON的BJSON格式的数据，其结构很松散，因而能够存储相对复杂的数据类型。支持的查询语言极其强大是Mongo一个最大的特点，Mongo的语法跟面向对象的查询语言有些类似，基本能够实现类似关系数据库单表查询中绝大多数的功能，同时还能对数据建立索引进行支持。它的特点是易使用、易部署、高性能，非常容易存储数据。

MongoDB的主要功能与特性如下。

- 面向集合存储，容易存储对象类型的数据；
- 支持动态查询；
- 模式自由；
- 支持完全索引，包含内部对象；
- 支持复制与故障恢复；
- 支持查询；
- 使用高效的二进制数据存储，包括大型对象（如视频等）；
- 自动处理碎片，以支持云计算层次的扩展性；
- 支持JAVA、RUBY、PHP、C++、PYTHON等多种语言；
- 文件存储格式为BSON（一种JSON的扩展）；
- 能够通过网络访问。

“面向集合”（Collection–Oriented）是指数据被分组存储于数据集中，被称作一个集合

① http://wenku.baidu.com/view/f996568683d049649b66588f.html.

（Collection）。在数据库中每一个集合都有着惟一的标识名，而且能够包含无穷数目的文档。集合的概念与关系型数据库（RDBMS）里的表相类似，不一样的是它无需定义任何模式。Nytro MegaRAID技术中的闪存高速缓存算法，能够快速识别数据库内大数据集中的热数据，提供一致的性能改进。

模式自由（schema-free），也就是存储在MongoDB数据库中的文件，关于它的任何结构定义用户并不需要知道①。假如知道的话，用户完全可以在同一个数据库里存储不同结构的文件。

文档主要以键值对的形式存储在集合中。键是字符串类型，用于惟一地标识一个文档，多种复杂的文件类型均可以是值，这种存储形式叫做BSON（Binary Serialized Document Format）。

MongoDB已经在多个站点部署，其主要场景如下。

- 网站实时数据处理。它非常适合实时地插入、更新与查询，并具备网站实时数据存储所需的复制及高度伸缩性。
- 缓存。由于性能很高，它适合作为信息基础设施的缓存层。在系统重启之后，由它搭建的持久化缓存层可以避免下层的数据源过载。
- 高伸缩性。非常适合由数十或数百台服务器组成的数据库，它的路线图中已经包含对MapReduce引擎的内置支持。

不适用的场景如下。

- 要求高度事务性的系统。
- 传统的商业智能应用。
- 复杂的跨文档（表）级联查询。

3.7 NewSQL数据库

NewSQL是对各种新的可扩展/高性能数据库的简称，这类数据库不仅具有NoSQL对海量数据的存储管理能力，还保持了传统数据库支持的ACID和SQL等特性。本节介绍了NewSQL数据库的相关概念，并详细介绍了MySQL Cluster技术和VoltDB等内容。

3.7.1 NewSQL数据库简介

人们曾经普遍认为传统数据库由于支持ACID和SQL等特性限制了其本身的扩展和处理海量数据的能力，因此尝试通过牺牲这些特性来提升对海量数据的存储管理能力。但是现在有些人持有不同的观念，他们认为制约系统性能的是锁机制、日志机制和缓冲区管理等，而不是支持ACID和SQL的特性。只要优化这些技术，关系型数据库系统在处理海量数据时也能获得很好的性能。

关系型数据库处理事务时对性能影响较大，需要优化的因素有以下5条。

（1）通信。应用程序通过ODBC或JDBC与DBMS进行通信是OLTP事务中的主要开销。

（2）日志。关系型数据库事务中对数据的修改需要记录到日志中，而日志则需要不断写到硬盘上来保证持久性，这种代价是昂贵的，而且降低了事务的性能。

① http://wenku.baidu.com/view/f996568683d049649b66588f.html.

（3）锁。事务中修改操作需要对数据进行加锁，这就需要在锁表中进行写操作，造成了一定的开销。

（4）闩。关系型数据库中一些数据结构，如B树、锁表、资源表等的共享影响了事务的性能。这些数据结构常常被多线程程序读取，所以需要短期锁即闩。

（5）缓冲区管理。关系型数据将数据组织成固定大小的页，在内存中磁盘页的缓冲管理会造成一定的开销。

为了解决上面的问题，一些新的数据库采用部分不同的设计，它取消了耗费资源的缓冲池，在内存中运行整个数据库。它还摒弃了单线程服务的锁机制，而通过使用冗余机器来实现复制和故障恢复，取代原有的昂贵的恢复操作。这种可扩展、高性能的SQL数据库被称为NewSQL。其中“New”用来表明与传统关系型数据库的区别。NewSQL是一个很宽泛的概念，它首先由451集团在一份报告中提出。NewSQL主要包括两类系统：拥有关系型数据库产品和服务，并将关系模型的好处带到分布式架构上；或者提高关系数据库的性能，使之达到不用考虑水平扩展问题的程度。

NewSQL能够提供SQL数据库的质量保证，也能提供NoSQL数据库的可扩展性。VoltDB是NewSQL的实现之一，其开发公司的CTO宣称，他们的系统使用NewSQL的方法处理事务的速度比传统的数据库系统快45倍。VoltDB可以扩展到39个机器上，在300个CPU内核中每分钟可处理1600万个事务，其所需的机器数比Hadoop集群要少很多。

NewSQL数据库的分类如图3.20所示。

NewSQL数据库
MySQL Cluster NuoDB SenseiDB
VoltDB Citrusleaf RethinkDB
Datomic GenieDB ScalArc
未开源 FathomDB Database.com
Xeround Amazon RDS SQL Azure

图3.20 NewSQL数据库的分类

3.7.2 MySQL Cluster

MySQL Cluster是一种允许在无共享系统中部署“内存中”数据库的Cluster技术。无共享体系结构的存在，使得系统能够使用廉价的硬件，而且系统对软硬件并没有特殊的要求。此外，由于每个组件都有自己的内存和磁盘，故不会存在单点故障。

MySQL Cluster由一组计算机构成，每台计算机上均运行着多个进程，包括MySQL服务器、管理服务器、NDB Cluster的数据节点以及专门的数据访问程序等。Cluster中这些组件的关系如图3.21所示。

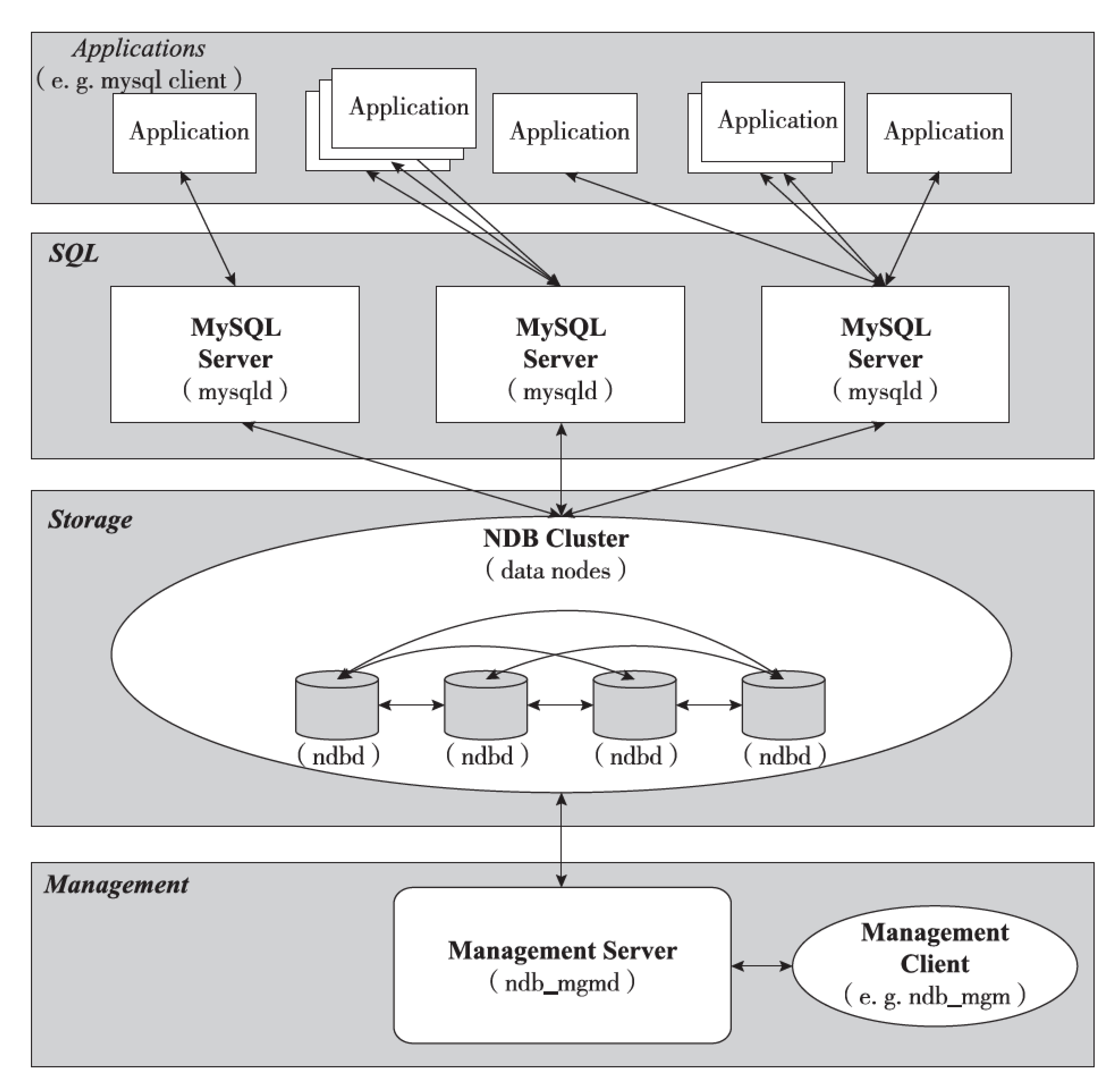


图3.21 Cluster中组件的关系

MySQL Cluster能够使用多种故障切换和负载平衡选项来配置NDB存储引擎。这里的“NDB”是一种“内存中”的存储引擎，它具有可用性高和数据一致性好的特点，但只有在Cluster级别上的存储引擎上实现这个技术最简单。

MySQL Cluster的Cluster有一部分可独立于MySQL服务器进行配置。在MySQL Cluster中，Cluster的各个部分都被视为1个节点。以下是几种节点的介绍①。

（1）管理（MGM）节点。这类节点的主要作用是管理MySQL Cluster内的其他节点，如提供配置数据、启动并停止节点等。由于它的功能特殊，因此，在启动其他节点之前应该首先启动这类节点。可以用命令“ndb_mgmd”来启动MGM节点。

（2）SQL节点。这类节点主要是用来访问Cluster的数据。对MySQL Cluster来说，客户端节点使用的是NDB Cluster存储引擎的传统MySQL服务器。通常情况下，SQL节点是使用命令“mysqld –ndbcluster”来启动的，或着也可以将“ndbcluster”添加到“my.cnf”后使用“mysqld”来启动。

（3）数据节点。这类节点是用来保存Cluster的数据。数据节点数目的多少是根据副本数目的多少来决定的，且这个数目是片段的倍数。比如，若有两个副本，而且每个副本均有两

① http://baike.so.com/doc/2286945.html.

个片段，那么就有4个数据节点。可以用命令“ndbd”来启动数据节点。

管理服务器（MGM节点）负责管理Cluster日志和Cluster的配置文件。Cluster中的每个节点从MGM节点检索配置数据，并请求确定MGM节点所在位置的方式。如果数据节点内出现新的事件，节点就会将关于这类事件的信息传输到MGM节点，然后，将这类信息写入到Cluster日志中。MySQL Cluster是MySQL适合于分布式计算环境的高实用、高冗余版本，MySQL Cluster采用了NDB Cluster存储引擎。在MySQL 5.0及以上的二进制版本和与最新的Linux版本兼容的RPM中就提供了该存储引擎。（注意，要想获得MySQL Cluster的功能，必须要安装mysql-server和mysql-max RPM）

所有的这些节点构成一个完整的MySQL集群体系。数据保存在“NDB存储服务器”的存储引擎中，表（结构）则保存在“MySQL服务器”中。应用程序通过“MySQL服务器”来访问这些数据表，集群管理服务器则通过管理工具(ndb_mgmd)来管理“NDB存储服务器”。

实践证明，通过将MySQL Cluster引入开放源码世界，MySQL为有需要的人提供了具有高可用性、高性能和可缩放性特点的Cluster数据管理。

此外，可以有任意数目的Cluster客户端进程或应用程序。它们主要分为以下两种类型[①]。

标准MySQL客户端：对MySQL Cluster来说，它们与标准的MySQL并没有区别。即，它们同样可以从用PHP、C、C++、Java、Python等编写的现有MySQL应用程序中访问MySQL Cluster。

管理客户端：这类客户端与管理服务器相连，并且这类客户端提供了启动和停止节点、启动和停止消息跟踪（仅调试版本）、显示节点版本和状态、启动和停止备份等命令。

3.7.3 VoltDB

VoltDB是一个运行于内存中的开源OLTP SQL数据库，能够保证事务的完整性。VoltDB是由Postgres和Ingres联合创始人Mike Stonebraker领导开发的一个下一代开源数据库管理系统。VoltDB能在现有的廉价服务器集群上处理每秒数百万次数据。

（1）VoltDB的特点

VoltDB具有如下特点。

- VoltDB大大降低了服务器资源的开销。
- 单节点每秒数据处理远远高于其他数据库管理系统。
- VoltDB可以使用SQL存取，支持传统数据库的 ACID模型。

（2）VoltDB的特征

为了获得最大化吞吐量，数据可以保存在内存中，这样可以有效地消除缓冲区管理。 VoltDB通过SQL引擎把数据分发给集群服务器的每个CPU处理。每个单线程分区自主执行，消除锁定和闩锁的需求。VoltDB可以简单地通过在集群中增加附加节点的方式实现性能的线性增加。

（3）VoltDB的组成部分

一个典型的 VoltDB 应用程序由以下文件组成。

- 一个项目定义文件 (project.xml)。其中包含哪些存储过程可用、数据库模式文件的位置、分区信息等信息。

① http://blog.csdn.net/chinalinuxzend/article/details/1768860

- 一个部署文件 (deployment.xml)。其中包含每个主机的站点数等信息。
- 数据库模式 (ddl.sql)。
- 源代码。

（4）VoltDB的使用要求

VoltDB 需要一个基于64位Linux 的操作系统（此要求也适用于Mac OSX 10.6），同时还需要安装Java开发工具包。可以使用 Eclipse来编辑源代码。

要将项目导入到 Eclipse 中，可打开 Eclipse，然后执行以下操作。

a. 选择File>New>Project。

b. 选择Java Project from Existing Ant Buildfile，然后单击Next。

c. 选择复选框Link to the build file in the file system。

d. 从刚刚安装示例应用程序的目录中选择build.xml作为Ant生成文件，然后选择Finish。

如果希望创建自己的应用程序，VoltDB 提供了一个工具来为用户生成框架项目，该项目用于生成本文中附带的应用程序的文件夹结构。可以用下面这个命令进行调用。

```
$ cd $HOME/voltdb-2.5/tools
$ ./generate app acme $HOME/Projects/app
```

运行上述命令并查看新创建的文件夹，将会看到该工具生成了一个框架项目。其中就包含了构建一个 VoltDB应用程序所需的文件。

（5）VoltDB的版本

目前VoltDB有两个版本：一个开源社区版本和一个付费企业版本。付费企业版本除了有开源社区版的所有功能外，还具备这些特点：计算机集群管理控制台、数据库当机恢复、在线数据库Schema修改、在线数据库节点重新加入以及命令日志等。

3.8 分布式缓存系统

在数据驱动的Web开发中，经常要重复从数据库中取出相同的数据，这种重复极大地增加了数据库负载。缓存是解决这个问题的好办法。现阶段比较流行的分布式缓存系统有Memcached、Ehcache、OSCache、JSC、Jmemcached和Tcache等。

下面简单介绍一下Memcached缓存技术。

1. Memcached简介

Memcached是由Danga Interactive开发的一个高性能分布式的内存对象缓存系统，在动态应用中可以帮助访问速度提升和使数据库负载减少[①]。通过在内存中对一个统一的巨大的hash表的维护，Memcached可以用来存储包括视频、图像、文件和数据库检索结果等多种格式的数据。Memcached缓存技术的特点是：它的缓存是一种分布式的，能够让不同主机上的多个用户同一时间进行访问。这打破了共享内存只能够在单机使用的局限，同时避免了在使用数据库做相似事情的时候，磁盘开销以及阻塞现象的发生。

① http://blog.csdn.net/webwalker/article/details/1584551

2. Memcached的使用

Memcached服务器端的安装（系统服务安装）步骤如下。

（1）下载文件：memcached 1.2.1 for Win32 binaries (Dec 23，2006)。

（2）将文件解压缩到c:\memcached目录。

（3）在命令行输入“c:\memcached\memcached.exe –d install”。

（4）在命令行输入“c:\memcached\memcached.exe –d start”。该命令将启动Memcached，默认监听端口是11211。通过执行memcached.exe －h命令能够查看其帮助。

.NET memcached client library的安装。

下载并安装文件：https://sourceforge.net/projects/memcacheddotnet/，里面有.net1.1与.net2.0两个版本。

Memcached的应用方法如下。

（1）将Commons.dll，ICSharpCode.SharpZipLib.dll，log4net.dll Memcached.ClientLibrary.dll等放入bin目录中。

（2）引用Memcached.ClientLibrary.dll。

（3）代码如下。

```
namespace Memcached.MemcachedBench
{
using System;
using System.Collections;
using Memcached.ClientLibrary;
public class MemcachedBench
{
[STAThread]
public static void Main(String[] args)
{
string[] serverlist = { "10.0.0.131:11211", "10.0.0.132:11211" };
//初始化池
SockIOPool pool = SockIOPool.GetInstance();
pool.SetServers(serverlist);
pool.InitConnections = 3;
pool.MinConnections = 3;
pool.MaxConnections = 5;
pool.SocketConnectTimeout = 1000;
pool.SocketTimeout = 3000;
pool.MaintenanceSleep = 30;
pool.Failover = true;
pool.Nagle = false;
pool.Initialize();
```

```
 // 获得客户端实例
 MemcachedClient mc = new MemcachedClient();
 mc.EnableCompression = false;
 Console.WriteLine("------------测  试-----------");
 mc.Set("test", "my value");  //存储数据到缓存服务器，这里将字符串"my value"缓存，
 key 是"test"
 if (mc.KeyExists("test"))   //测试缓存存在key为test的项目
{
 Console.WriteLine("test is Exists");
 Console.WriteLine(mc.Get("test").ToString());  //在缓存中获取key为test的项目
}
 else
{
 Console.WriteLine("test not Exists");
}
 Console.ReadLine();
 mc.Delete("test");  //移除缓存中key为test的项目
 if (mc.KeyExists("test"))
{
 Console.WriteLine("test is Exists");
 Console.WriteLine(mc.Get("test").ToString());
}
 else
{
Console.WriteLine("test not Exists");
}
Console.ReadLine();
SockIOPool.GetInstance().Shutdown();  //关闭池， 关闭sockets
}
}
}
```

（4）运行结果如图3.22所示。[①]

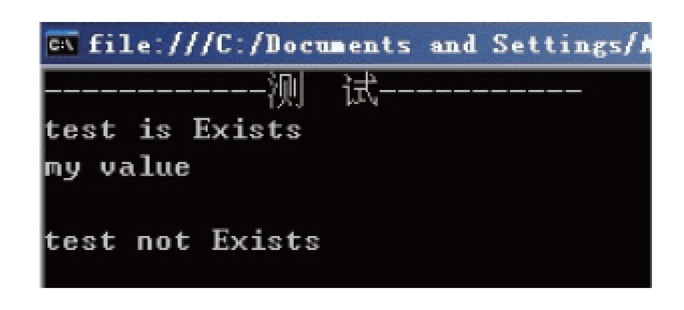


图3.22　运行结果

① http://blog.csdn.net/webwalker/article/details/1584551

3. Memcached缓存系统的优缺点

提高了访问获取数据的速度，也可用于加速Web应用。但与此同时Memcached的缓存系统没有特殊的安全机制，需要自己控制安全。

适用场景：缓存对性能影响较大的数据、缓存变动不是很频繁的数据。

3.9　练习

1. 什么是NoSQL?
2. 什么是关系型数据库?
3. 关系型数据库的优点与缺陷是什么?
4. NoSQL的优点与缺陷是什么?
5. 关系型数据库与NoSQL之间的区别与联系是什么?
6. 什么是NoSQL的三大基石?
7. CAP的三个特性是什么?
8. 分布式系统为什么不能同时满足CAP理论的三个特性呢?
9. ACID与BASE的区别是什么?
10. NoSQL的典型分类是什么?
11. 什么是键值数据库?
12. Bigtable的特点是什么?它属于NoSQL四大分类中的哪一类?
13. 什么是HBase?它的特点有什么?
14. 什么是图存数据库?
15. 典型的文档数据库有哪些?请简单描述一下这些文档数据库的特点与工作原理或场景。
16. 什么是NewSQL数据库?
17. 简要说明什么是分布式缓存系统。

参考文献

[1] 陆嘉恒. 大数据挑战与NoSQL数据库技术[M]. 北京：电子工业出版社，2013.
[2] 佐佐木达也. NoSQL数据库入门[M]. 北京：人民邮电出版社，2012.
[3] 塞得拉吉，福勒. NoSQL精髓[M]. 北京：机械工业出版社，2013.
[4] Shashank Tiwari. 深入NoSQL[M]. 北京：人民邮电出版社，2012.
[5] 张华强. 关系型数据库与NoSQL数据库[J]. 电脑技术与知识，2011.

第4章 Hadoop和MapReduce

大数据是当今整个世界的趋势。全球每时每刻都有大量的数据生成，当人们在给别人发微信、分享视频、上传拍照、更新社交状态、转发他人微博或者论坛上发帖的时候，这些操作都会使得机器设备产生和保留数据。面对这样迅猛增长的数据，世界上一些处于市场领导地位的互联网公司，如谷歌、雅虎、BAT（百度、阿里巴巴、腾讯）等感受到了巨大的压力。这些公司的人员需要分析处理大如TB 、PB甚至EB级别的数据量，来得到一些有价值的信息。比如哪些产品的销量比较大，哪些新闻的点击率高，哪些广告更能获得用户的喜爱。但目前出现的问题是，传统的工具越来越难以对如此规模的数据集进行存储、计算和分析，这也是大数据时代的一个很明显的特征。所以Hadoop出现了，它正是为了解决大数据时代方便大量数据存储和处理而研发出来的框架，是大数据最重要的技术。

4.1 Hadoop简介

什么是Hadoop呢？Hadoop 是由Apache基金会开发出来的一个开源的软件框架，简单地说，Hadoop是一个分布式系统和并行执行环境，便于存储和处理大规模数据的开源软件平台图4.1是Hadoop的图标。

图4.1 Hadoop的图标

Hadoop有四个主要的特点，分别是：扩展能力强、成本低、高效率和可靠。

- 扩展能力强（Scalable）：能轻易地存储和处理千兆字节（PB）的大数据。
- 成本低（Economical）：通过使用数千个普通机器组成的服务器群来进行数据分发以及进行处理等工作、代替了性能高的昂贵机器。
- 高效率（Efficient）：通过分工合作的方式将数据分发到各个机器上，由Hadoop并行地处理它们，这使得整个处理过程变得非常的快速。
- 可靠（Reliable）：Hadoop具有数据自动备份和维护机制，能有效解决任务失败和数据丢失后的恢复和重新部署计算任务问题。

Hadoop目前具有很多项目，包括HBase、Pig、Common、Avro、ZooKeeper，以及被称为核心的Hadoop分布式文件系统HDFS和MapReduce等，项目组成如图4.2所示。

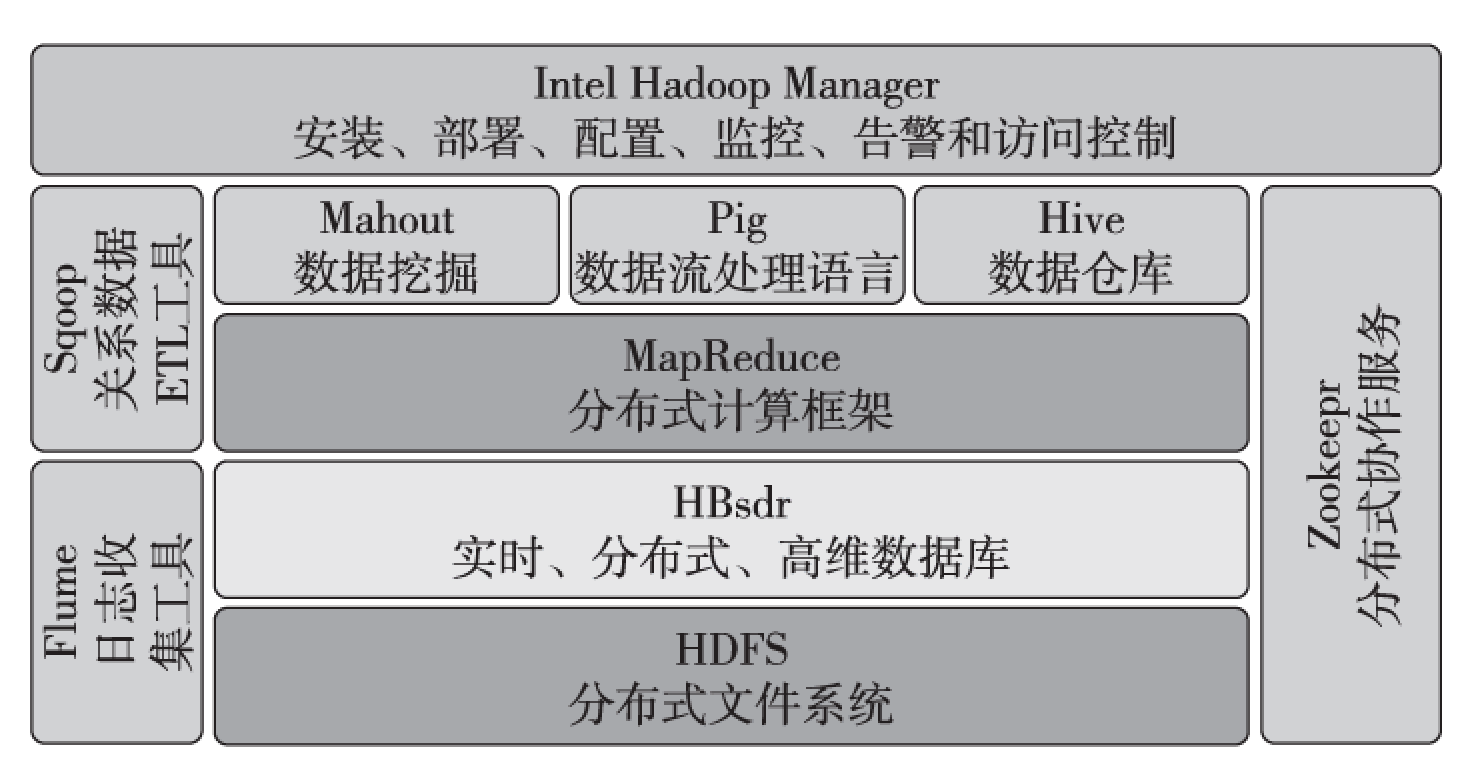


图4.2　Hadoop的项目结构图

4.2　Hadoop的体系结构

正如前面所提到的，HDFS和MapReduce是Hadoop的两大核心，下面主要详细介绍HDFS和MapReduce。

4.2.1　HDFS的体系结构

HDFS是Hadoop分布式文件系统（Hadoop Distributed File System）的缩写。它是Hadoop 体系中数据存储和管理部分的关键。HDFS有高度容错的功能，能检查出哪些部件发生了故障并对这些部件进行处理，因此它可以设计部署在价格低廉的硬件上。HDFS 简化了文件的一致性模型，采用流式数据访问，实现了高数据吞吐量，更适用于具有大规模数据集的应用程序。HDFS降低了对可移植操作系统接口（POSIX，Portable Operating System Interface）的要求，实现了对文件系统中的数据的流式访问。HDFS是Hadoop的基础架构之一。一个HDFS集群有两类节点，它们以管理者—工作者模式进行运作，即由一个NameNode（管理者）和若干个DataNode（工作者）构成。此外还有与这两个角色之间作为沟通桥梁的Client（客户端）。图4.3是HDFS体系结构图。

图4.3　HDFS体系结构图

接下来对其节点进行详细介绍。

1. NameNode

NameNode在系统中充当“管理者”的角色，主要负责文件系统的NameSpace（命名空间）的管理，以及管理客户端对文件访问的一些常用的操作，比如文件的打开、关闭、保存或是移动文件等，与此同时它还负责把数据块映射至对应的DataNode上。此外，NameNode

还负责维护文件系统树，以及这棵系统树内的所有文件和目录。当它保存在本地磁盘上时，有命名空间镜像文件和编辑日志文件这两种文件形式。NameNode主要负责去记录数据块的信息，只是它并不具备永久保存的功能，因为当系统启动时，NameNode上的信息会再次被建立，这是由数据节点完成的。也就是说，脱离了NameNode，整个文件系统就会陷入瘫痪。所以，对NameNode实现容错功能是非常重要的。Hadoop特地为此提供了两种机制：第一种机制是对写入的文件进行备份；第二种机制是运行一个辅助的NameNode（即Secondary NameNode）用来时刻检测HDFS的集群状态，以减少数据丢失的几率。

2. DataNode

DataNode在文件系统中作为“工作者”的角色，负责存储并检索数据块，每个文件都被分为很多个数据块，这些数据块存储在DataNode上。它也负责完成客户端访问文件的读写操作，定期向NameNode发送它们所存储的数据块的列表，并在NameNode的统一调度下开展工作，如数据块的创建、移动、复制和删除等。

3. Client

Client代表用户通过与NameNode和DataNode来交互访问整个文件系统。因为Client提供了一些文件系统接口，所以在编写程序的时候，不用知道DataNode和NameNode的内部详细情况，就可以编写程序进行操作，实现所需的功能。

4.2.2 MapReduce的体系结构

MapReduce是一种针对大规模数据集（通常都是大于1TB）并行计算而设计的编程模型。MapReduce、Common、HDFS属于Hadoop发展初期的三个组件。在Hadoop体系中，可以利用MapReduce模型对任务进行分配，使分配后的任务可在上千台商用机器组成的集群上进行并行计算，加上本身具有的高容错的特性，实现了Hadoop对任务的并行处理功能。

MapReduce主要分为Map（映射）和Reduce（归约）两个步骤，其中Map负责对数据集上的独立元素进行特定的操作，生成Key-Value（键-值）对形式的中间结果；Reduce则负责对中间结果中相同“键”的所有“值”进行合并规约，并且得到最后的结果。MapReduce这样的功能划分和处理步骤，在大量机器组成的分布式并行操作的环境里进行数据计算处理显得尤为合适。因此，开发人员需要做的只是编写Map和Reduce两个函数，即可完成一个简单的分布式程序的设计工作。这从某种程度上，为程序员提供了很大的便利——即使他们不熟悉分布式并行计算系统编程的具体细节，也可以顺利地在分布式的系统上运行其编写的代码。

简单来说，HDFS是负责分布式存储数据的，MapReduce是负责分布式数据计算的，因此这也就是为什么Hadoop作为很多分布式计算的平台的原因了。

4.2.3 其他组件

简单介绍一下其他组件。

1. HBase

HBase 是一个动态的数据库，它针对结构化的数据，具有可伸缩、高可靠、高性能、

分布式和面向列的特点。与传统关系数据库不同的是，HBase 采用了增强的稀疏排序映射表（Key/Value）这样一种谷歌BigTable 的数据模型。其中，HBase的键由三个部分组成，分别是行关键字、列关键字和时间戳。同时，HBase也提供了一些对大数据的操作访问，比如随机和实时的读写等。另外，也能用MapReduce去处理在HBase 中存储的数据，这样一来就能将数据的存储和并行计算进行有机地结合。

2. Pig

Pig是一个基于Hadoop的平台，主要负责分析和评估大型数据集。它使得利用Hadoop来进行数据分析的条件变得更加简单，为此专门设计了一个高级的、面向领域为特征的抽象语言，即Pig Latin。有了Pig Latin，与数据分析相关的工作人员能够对结构多样且相互之间有联系的数据分析任务进行编码，将数据分析工作转化成基于Pig操作的数据流脚本，并且再将该数据流脚本编码为MapReduce工作链，使其能在Hadoop的环境里运行。与Hive相同的地方在于，Pig也能够使分析和评估大型数据变得更加简单。

3. Common

Common（Hadoop 0.20之前称为Core）是一种能为Hadoop结构中的其他项目提供常用工具的项目，这些工具大致有：远程调用、抽象文件系统、系统配置工具和序列化机制等。这些工具提供了一些基础服务，帮助通用硬件完成云计算环境的搭建，同时包含了一些软件开发时使用的API（应用程序接口）。

4. Avro

Avro是一个数据序列化的系统。Avro能够通过转换，将具有结构化或者非结构化的复杂格式类型的数据或对象变成其他格式。这些格式最大的特点是容易保存和传送，这与别序列化机制很像。Avro的主要设计目的是希望能够支持一些以数据密集型为特征的程序，这样能有利于进行大规模的数据存储与交换。此外，它还提供了许多强大的功能，包括保存永久性数据文件集合、集成简单动态语言、提供快速的可压缩的二进制数据格式、远程过程调用以及多样的数据结构类型等。

5. ZooKeeper

ZooKeeper是一个分布式框架，主要用于处理分布式计算中出现的一致性问题。所谓一致性，是指事务的基本特征或特性相同，其他特性或特征相类似，比如保持数据格式的统一。在分布式系统中，怎样对某个结果的某些特征保持一致，是非常值得考虑的问题。所以，ZooKeeper提供了很多解决分布式应用中出现的数据操作问题的服务方案和管理计划，比如命名一致、状态协同服务以及集群、分布式程序项目配置的管理等。ZooKeeper良好地协调了Hadoop中其他项目的工作并且也逐渐成为它们的主要部件，因此它的重要性也日益增加。

4.2.4 Hadoop的I/O操作

对于数据的输入和输出，Hadoop设计了一系列原子操作，其中的一些诸如数据完整性保持和数据压缩之类的重要技术，在用于操作海量数据时，显得非常重要。此外由其他的

Hadoop工具或者应用程序接口所组成的功能模块也能应用于开发分布式系统和结构，比如数据的序列化机制以及在盘（on-disk）数据结构。

1. 数据完整性

在传输过程中要保证数据不发生缺失和顺序颠倒等问题（即数据完整性）是非常重要的。通常，确保数据完整性的技术有以下三种。

- CRC-32循环冗余校验技术
- 奇偶校验技术
- ECC校验纠错技术

因为Hadoop采用HDFS作为其默认的文件系统，所以需要讨论两部分的数据完整性：HDFS的数据完整性和本地文件系统的数据完整性。

（1）HDFS的数据完整性

HDFS会在数据传输的以下三个阶段检验校验和。

① DataNode获得数据后，保存数据前

一般说来，DataNode所得到的数据有两种情况：一是使用者利用客户端完成上传数据的操作；二是DataNode在其他的数据节点上获得的内容。通常客户端在上传数据时会将数据传到一个专用的数据节点管道中，在该管道的最后一个数据节点去检验校验和。

② 客户端访问DataNode上的数据内容时

此时会进行校验和校对，方法是将DataNode上存储的校验和进行检测。

③ DataNode的后台守护线程对其进行定期检查时

DataNode会在后台运行一个DataBlockScanner（数据块检测程序），隔一段时间就会检测保存在DataNode上的所有数据块，以防止一些由物理存储介质（如硬盘）中因为位衰减所导致的数据丢失。

（2）本地文件系统的数据完整性

Hadoop的本地文件系统负责客户端的校验工作，因此客户端也需要负责本地文件系统的数据完整性。详细的操作是，每当Hadoop建立一个文件a时，Hadoop就会同一时刻在同一个文件夹地址下创建a的隐藏文件a.crc，这个隐藏文件中包含有文件a的数据校验和。根据具体数据文件的字节大小，每隔512个字节便会产生4个字节的校验和，如果要改变每个校验和所针对的文件的大小，开发人员可以在core-default.xml中通过改变io.bytes.per.checksum的值来实现。

此外，数据恢复策略也是值得开发人员注意的地方。当进行数据读取的操作时，如果发现某个数据块即将失效，那么作为与读取操作相关的部分，DataNode和NameNode都会采取复制完整备份的方式来恢复此数据块，并且在恢复成功的基础上设置标签，以防止此数据块被其他的角色进行重复恢复的操作。

2. 数据的压缩

文件压缩对于任何大容量的分布式存储而言意义重大，因为文件压缩具有两个优点，减小文件存放所需要的空间；提高文件的传送速度。

许多压缩格式和压缩算法在Hadoop中同样适用，只是不同的算法有各自不同的特点，如表4.1所示。

表4.1 Hadoop支持的压缩格式及特点

压缩格式	工具	算法	文件扩展名	多文件	可分割性
DEFLATE	无	DEFLATE	.deflate	不	不
Gzip	gzip	DEFLATE	.gz	不	不
Bzip2	bzip2	bzip2	.bz2	不	是
LZO	lzop	LZO	.lzo	不	不

（1）压缩与输入分割

通常需要对那些由MapReduce处理的数据进行压缩，而在压缩时，用户可以选择不同的压缩格式。对于比较大的压缩包来说，如果该压缩格式支持分割功能的话，对数据存储是非常有利的。例如，一个gzip格式的压缩文件，它压缩后有1GB，通过HDFS将这个压缩文件分为16块。然而想再对每一个块进行细分操作则是无法完成的，因为gzip格式的压缩文件不支持分割功能。因此，如果要读取它的MapReduce分割文件，会导致读取到16块大小相同的文件的情况，导致运行的时间变长。因此，总的原则是，需要经过测试，才可以决定某种压缩格式是否具有分割机制。如同以上所提到的，对于占用空间大的文件应该选择支持分割的压缩形式，如bzip2。

（2）在MapReduce程序中使用压缩

在MapReduce中使用压缩非常简单，只需要在它进行Job配置时，设定好conf的内容就可以了。

设置Map处理后压缩数据的代码如下。

```
JobConf conf = new Jobconf();
conf.setBoolean("mapred.compress.map.output",true);
```

对于一般情况来说，压缩是有利于文件存储和传输的，无论是对最终结果进行压缩，亦或是对Map处理后的中间结果进行压缩。

3. 序列化

（1）什么是Hadoop的序列化

序列化（serialization）指的是将结构化对象转为字节流的过程，这样有利于通过互联网进行传送或写入永久存储的介质中。反序列化则是序列化的逆过程，指的是将字节流转化为结构化对象的操作。

序列化在两大领域中经常出现：一个是进程间通信，另一个是数据持久性存储。

Hadoop使用远程过程调用来达到进程间通信的目的。通常情况下，远程过程调用的序列化机制有如下这些特征。

- 紧凑（Compact）：即易于通过网络进行传送，并且充分利用网络传输速率资源和存储资源。
- 快速（Fast）：即Hadoop的序列化和反序列化的能力较强，效果较好。
- 扩展性（Extensible）：即这些进程间通信所用到的协议要能有适应变化的能力，能够适应新的要求。
- 互操作性（Interoperable）：即通过设计一种共同遵守的格式的方式，使得客户端及

服务器端的操作不依赖于某种固定的语言。

序列化在Hadoop中是很重要的一种操作，因为用户要通过序列化的操作来存储文件和传送数据。对于具体的操作方式，Hadoop中并没有采用Java所具有的序列化机制，而是自己重新写了一个序列化机制Writable。Writable具有紧凑、快速的特点，所以使用起来更加方便。

（2）Writable接口

作为Hadoop的核心，Hadoop通过Writable定义了Hadoop中一些基本的数据类型和基本操作。通常，不管是上传下载数据，还是运行相关的MapReduce程序，Writable类的适用范围非常广。Writable类的结构如图4.4所示。

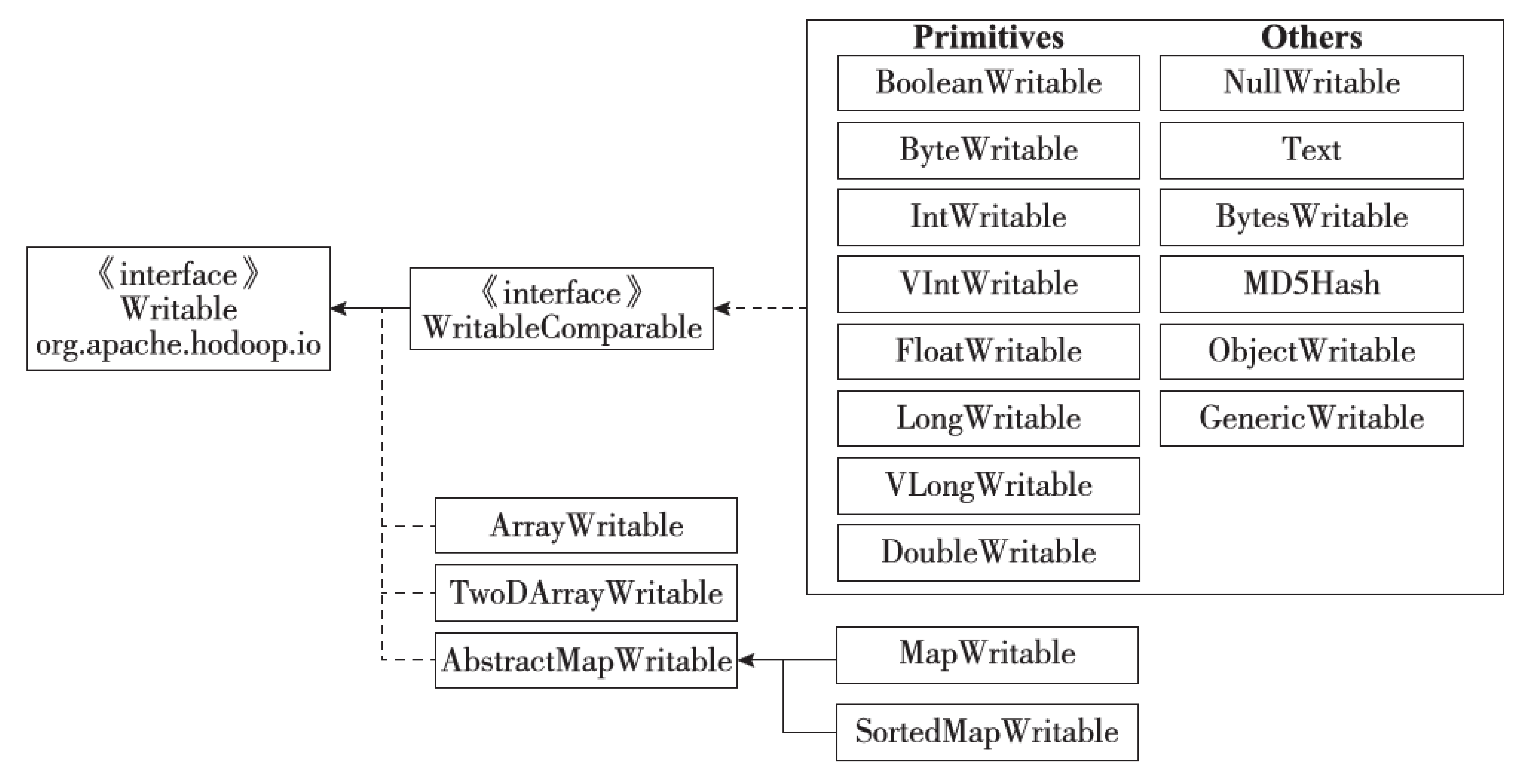


图4.4　Writable类的结构图

① Hadoop的比较器

WritableComparable是Hadoop中十分关键的接口类。它是一个比较器（一种能够比较不同的数据项之间的大小关系的装置）。

② Writable类中的数据类型

主要是Java的基本数据类型，例如boolean、byte、int、float等。

③ 其他类

NullWritable：这是一个占位符，其序列化长度为0。

BytesWritable和ByteWritable：BytesWritable是一个二进制数据数组的封装；ByteWritable是一个二进制数据封装。

Text：Hadoop是对string类型的重写，它使用的标准是UTF-8编码。

ObjectWritable：这是一种多类型的封装。

ArrayWritable和TwoDArrayWritable：分别是针对数组和二维数组的构建的数据类型。

MapWritable和SortedMapWritable：分别是java.util.Map()和java.util.SortedMap()的实现。

CompressedWritable：保存压缩数据的数据结构。

GenericWritable：通用的数据封装类型。

VersionedWritable：一个抽象的版本检查类。

4. 基于文件的数据结构

Hadoop定义了一些基于文件的数据结构，用来满足MapReduce编程框架的要求，其中核心是SequenceFile和MapFile这两种类型。Map过程所输出的中间结果就是由这两种数据类型表示的，其中MapFile是经过了排序并且具有索引功能的SequenceFile。下面分别进行介绍。

① SequenceFile

SequeceFile是由Hadoop 的应用程序接口所提供的一种二进制文件类型。这种二进制文件直接将键–值对进行序列化并保存到文件中。它所记录的是键–值对的清单，由于它是序列化之后的二进制文件，因此用户不能直接查看结果，但能够通过输入命令来查看文件的内容。

在sequenceFile中有以下三种不同类型的键–值对。

- 未压缩的key/value对。
- 记录压缩的key/value对（只有value被压缩）。
- Block压缩的key/value对。

SequenceFile类具有如下特点。

- 可压缩性：该类型文件能够被压缩，而且能定制为基于Record或Block压缩（其中Block级压缩性能较优）。
- 本地化任务支持：因为文件支持被切割，所以执行MapReduce任务时数据的本地化的情况应该是非常好的。
- 难度低：因为Hadoop框架提供了应用程序接口，所以很容易就能够修改业务逻辑。

② MapFile

MapFile的使用与SequenceFile很相像，但不同的是，SequenceFile生成的是一个文件，而这个MapFile生成的是一个文件夹。正如前面所提及的，它是经过排序并且具有索引功能的SequenceFile，因此用户能够通过键值进行检索。

MapFile由包含两部分，分别是data和index。index指的是文件的数据索引，里面含有每个记录的键值以及该记录在文件中的偏移量大小。在访问MapFile 的时候，首先在内存中加载索引文件，并且利用索引的对应关系能够立刻找到指定的记录文件在哪里。所以， MapFile的检索效率比SequenceFile高，但其劣势是必须要牺牲掉一部分的内存来保存它的 index值。

③ 其他文件类型

ArrayFile用于存储从Integer到值的对应关系。

SetFile与Java中的set作用相似，只是一个键的集合，而没有任何值。

BloomMapFile是具有过滤功能的MapFile。

4.2.5　Hadoop与分布式开发

什么是分布式系统？分布式系统主要是指能够进行分布式处理的软件系统，即多台机器在网络连通的情况下能一起工作的软件系统。分布式系统主要有：分布式的操作系统、文件系统、程序设计语言及其编译（解释）系统和数据库系统等。Hadoop属于文件系统，具有部分数据库的功能；HDFS负责数据的存储和管理；而MapReduce这个编程模型为用户编写Hadoop并行应用程序的运行提供了极大的便利。接下来主要谈谈在Hadoop基础上进行分布式

并行应用程序开发的相关知识。

开发并行应用程序，MapReduce框架必不可少。前文也提及了，其主要原理是利用两个函数Map和Reduce来实现并行程序，通过一批输入的键值对的集合来得到一批更简练的键值对的集合作为结果。详细步骤是：Map得到一个用户输入的键值对，通过处理产生中间的键值对，进行合并之后，再用一个迭代器把中间值传递给Reduce，最后得到一个相对较小的值的集合。通常每次使用Reduce时输出的值最多只有1个。

图4.5是MapReduce的数据流图，这个过程简单说来就是将整个大数据集合划分成多个小数据集，并且分配给集群中的若干个普通机器去处理，得到一些中间结果，接着进行排序和合并，获得最后的结果。具有MapReduce结构的并行程序内有三个主要的函数：Map、Reduce和main，但在这个结构中，需要用户编写的仅仅是Map和Reduce两个函数而已。

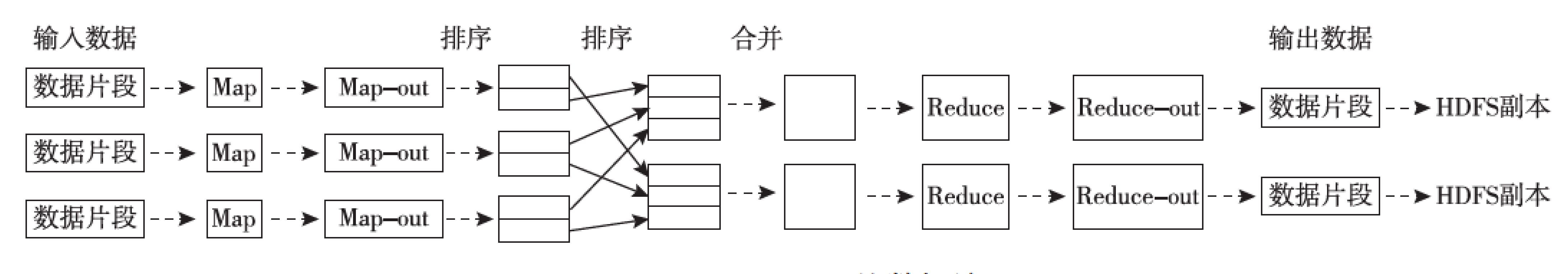


图4.5 MapReduce的数据流图

4.3 Hadoop的安装与配置

本节主要讲述如何在两个常用的操作系统：Windows和Linux上，安装和配置Hadoop的具体过程。

4.3.1 在Windows上安装与配置Hadoop

在Windows上运行Hadoop比较复杂，因为首先必须安装Cygwin，目的是用于模拟Linux环境，然后才能安装Hadoop。以下是详细的步骤①。

1. 安装JDK

JDK（Java Development Kit）是开发和运行Java程序的工具包。不建议只安装JRE，而建议直接安装JDK，因为安装JDK时会同时安装JRE。再者，MapReduce程序的编写和Hadoop 的编译都依赖于JDK，仅有JRE是不够的。JDK可在以下地址下载：

http://java.sun.com/java se/downloads/index.jsp

选择Java SE 即可。

2. 安装Cygwin

Cygwin是一个在Windows平台上运行的类UNIX模拟环境。在安装Cygwin前，需先下载Cygwin安装程序setup.exe。Cygwin 安装程序下载地址如下：

http://www.cygwin.com/setup.exe

当然也可以从http://www.cygwin.cn/setup.exe下载Cygwin 安装程序，不过如果在安装过程

① http://www.docin.com/p-64136032.html.

中，遇到如图4.6所示的错误，则只能从http://www.cygwin.com/setup.exe 下载。本书安装程序下载的是Cygwin 1.7.1版本。

图4.6　错误提示信息

Cygwin 安装程序setup.exe对存放的目录无要求。当setup.exe程序下载成功后，运行setup.exe，会弹出如图4.7所示的对话框。

在图4.7所示的对话框中，直接单击“下一步”按钮，进入如图4.8所示的对话框。

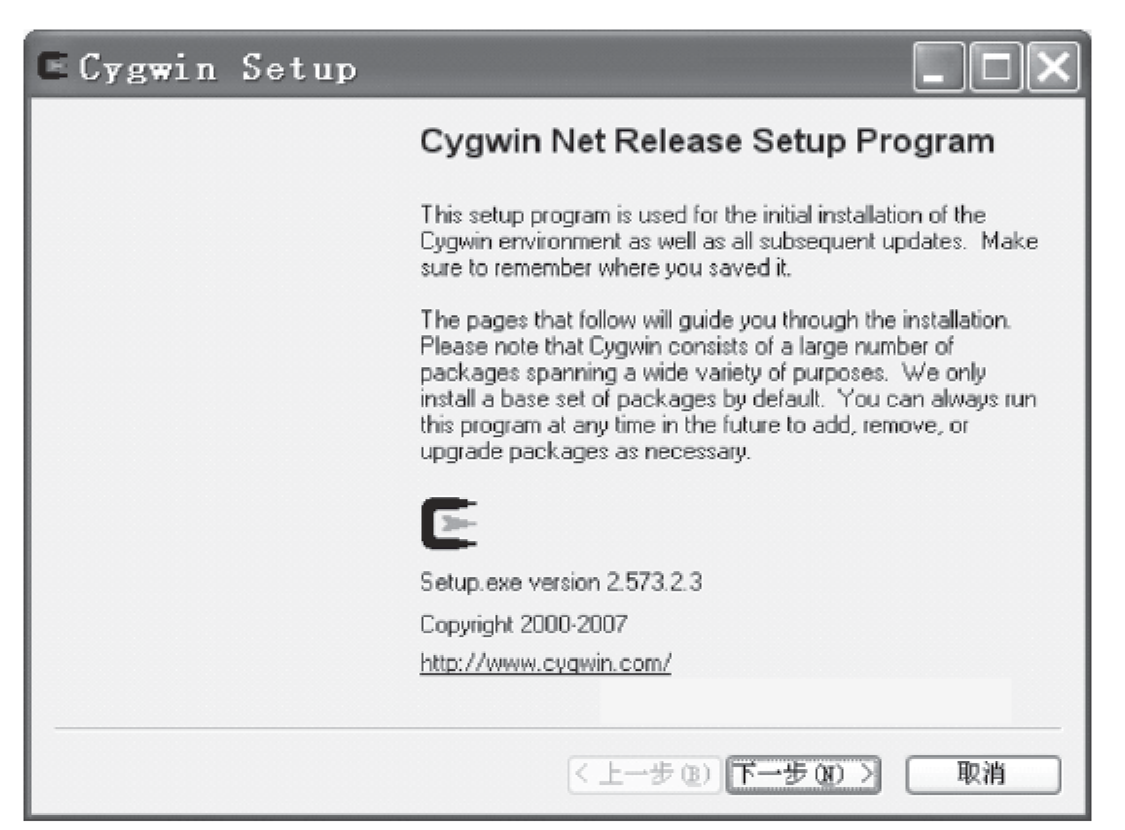


图4.7　Cygwin Setup对话框

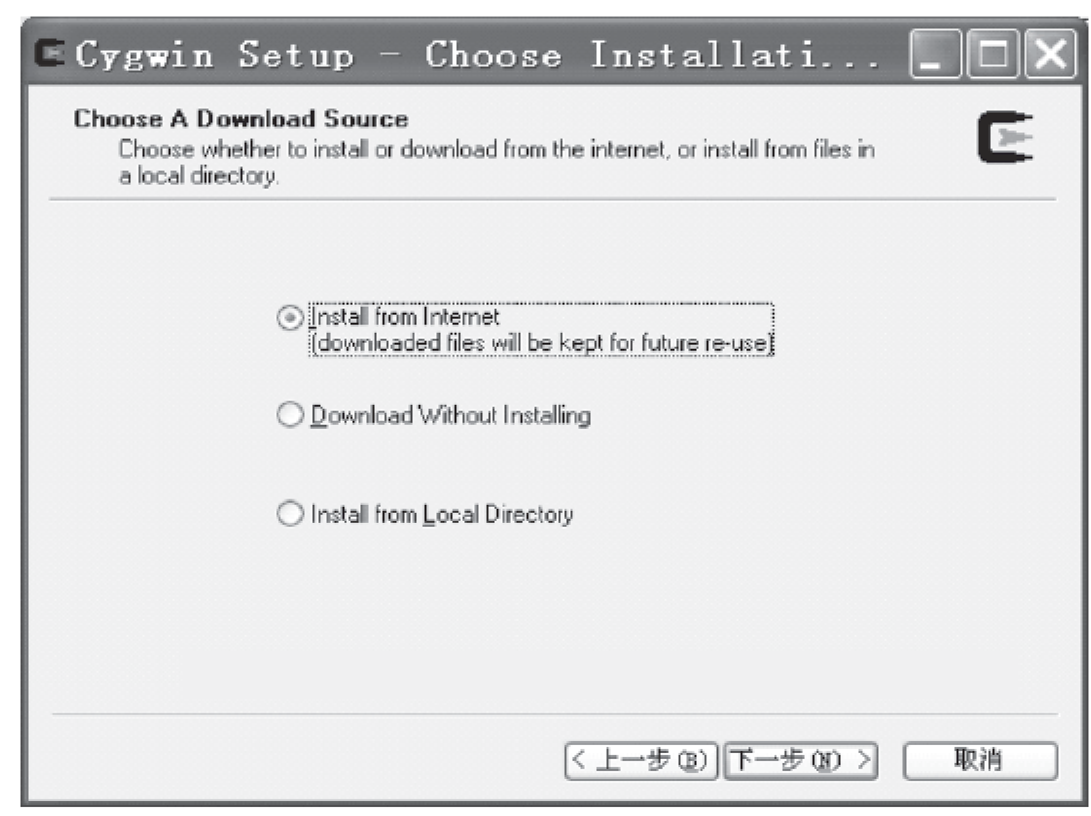


图4.8　选择“Install from Internet”

在图4.8所示的对话框中，选择“Install from Internet”复选框，然后单击“下一步”按钮，进入如图4.9所示的对话框。

在图4.9所示的对话框中，设置Cygwin 的安装目录，在“Install For”下选择“All Users”单选按钮，在“Default Text File Type”下选择“Unix/binary”单选按钮，然后单击“下一步”按钮，进入如图4.10所示对话框。

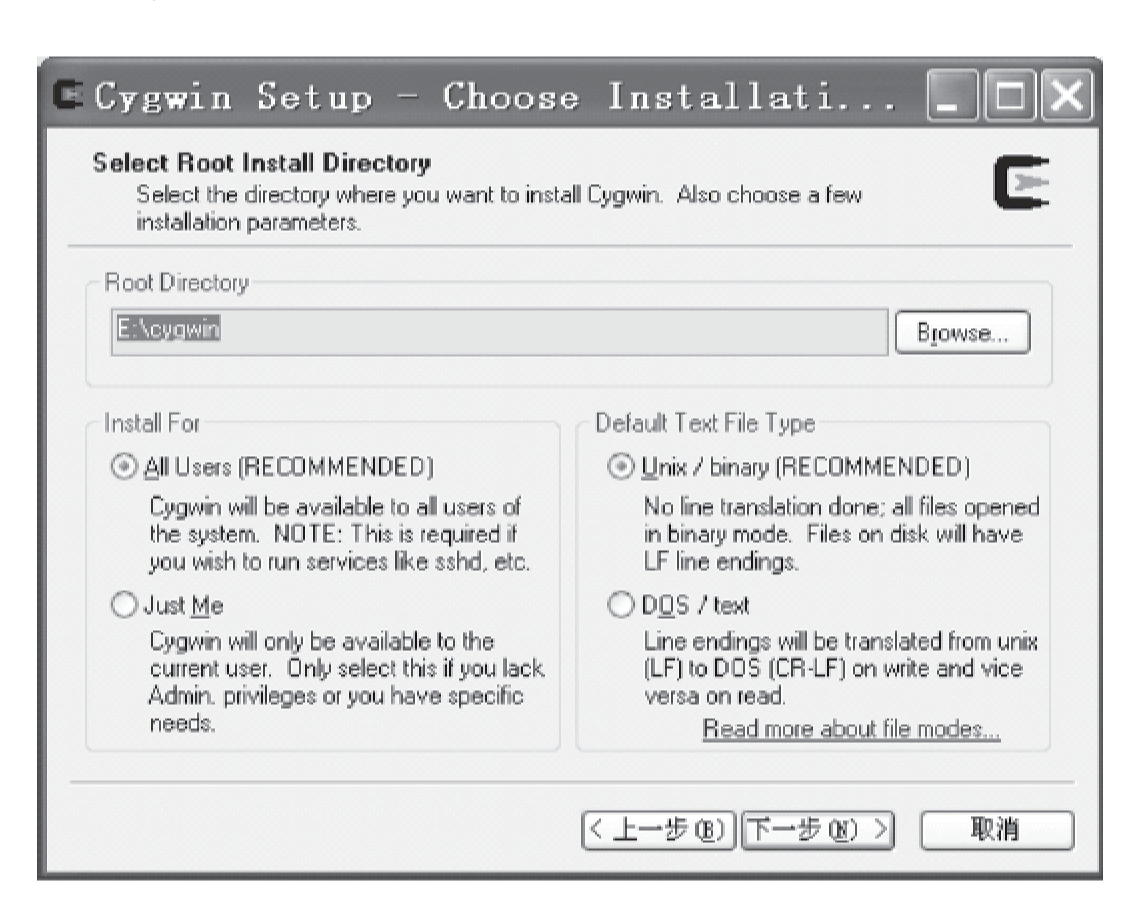


图4.9　设置Cygwin的安装目录

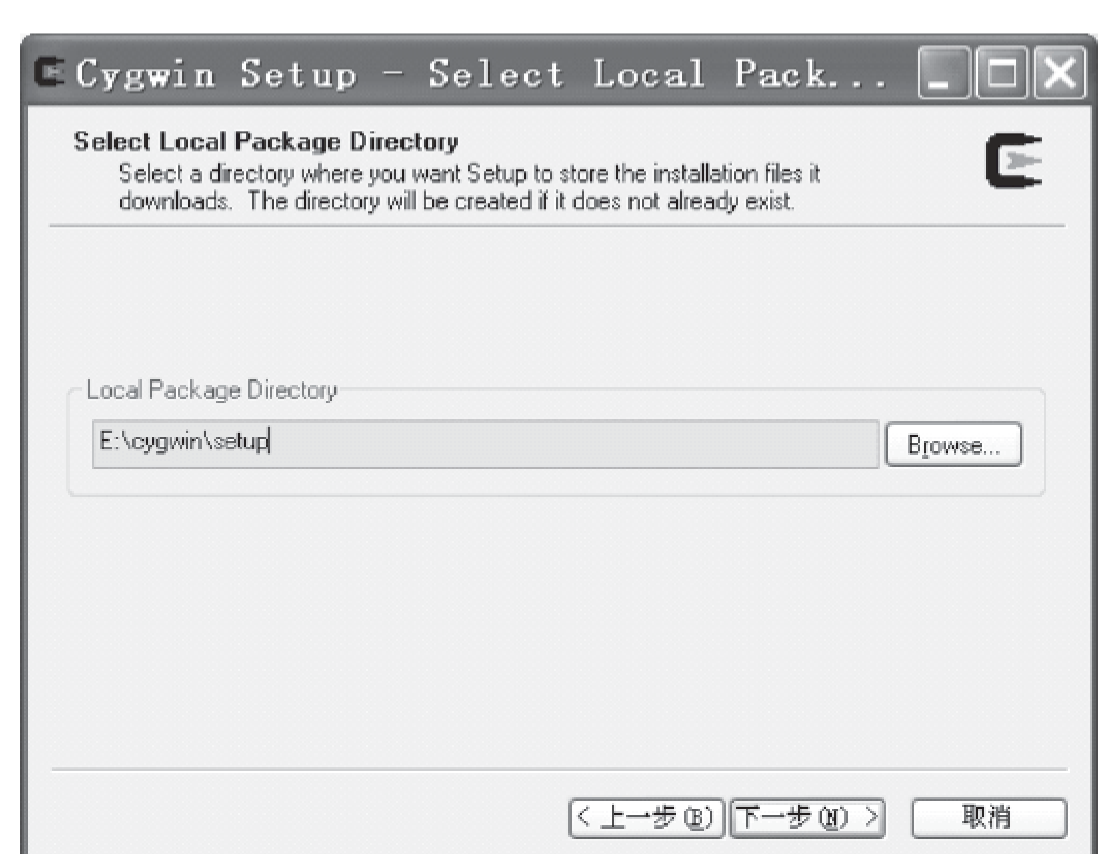


图4.10　设置安装包存放目录

在图4.10所示的对话框中，设置Cygwin 安装包存放目录，接着单击“下一步”按钮，就会出现图4.11所示的对话框。

在图4.11所示的对话框中， 单击选择“Direct Connection”单选按钮，然后单击“下一步”按钮，即出现如图4.12所示对话框。

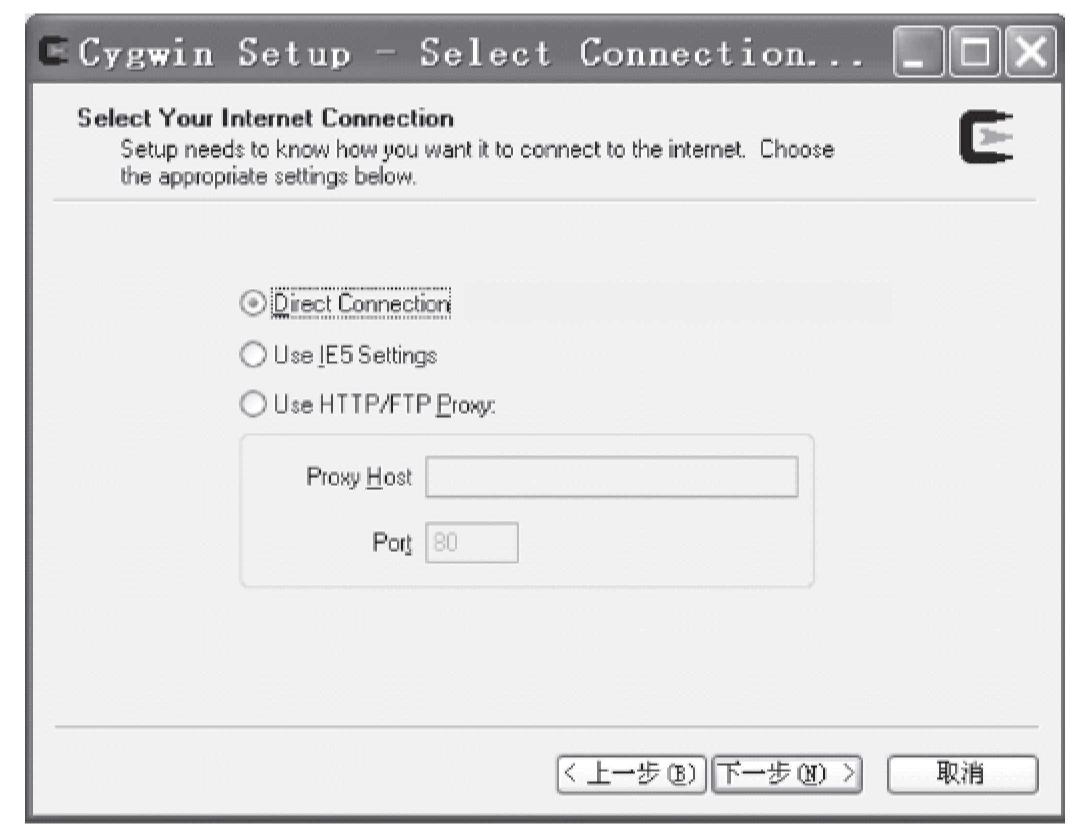


图4.11　选择“Direct Connection”

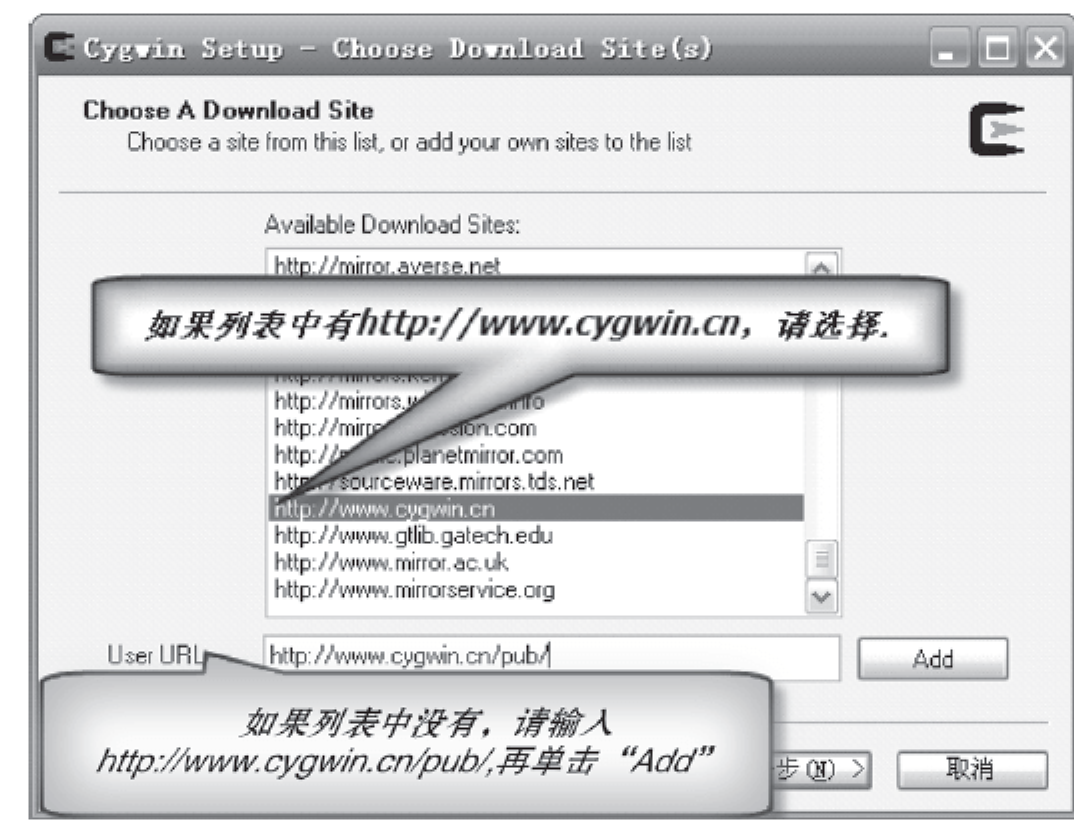


图4.12　设置路径

于图4.12所示的对话框中，单击“下一步”按钮，将进入如图4.13所示的对话框。

在图4.13所示的对话框过程中，可能会弹出如图4.14所示的“Setup Alert”提示框，直接单击“确定”按钮即可。注意：1.7.1之前的版本，不一定会弹出这个对话框。

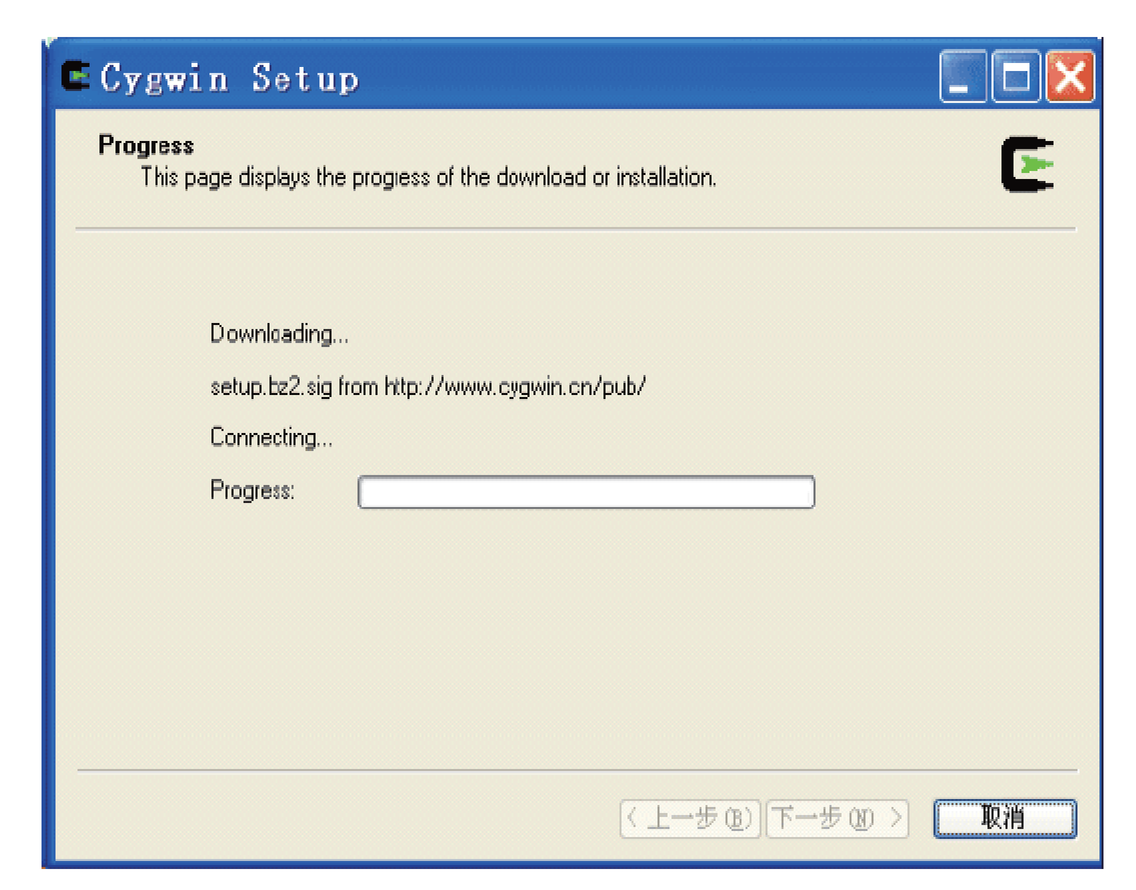


图4.13　在Windows上安装和配置Hadoop

图4.14　Setup Alert框

进入“Select Packages”对话框后，必须保证“Net Category”下的“OpenSSL”被安装，如图4.15所示。

如果还打算在eclipse 上编译Hadoop，则还必须安装“Base Category”下的“sed”，如图4.16所示。

另外，还建议将“Editors Category”下的“vim”也安装进来，以方便在Cygwin上直接修改配置文件。“Devel Category”下的“subversion”建议安装，如图4.17所示。

当实现上述操作后，单击“Select Packages”对话框中“下一步”按钮，即可进入Cygwin安装包下载过程，如图4.18所示。

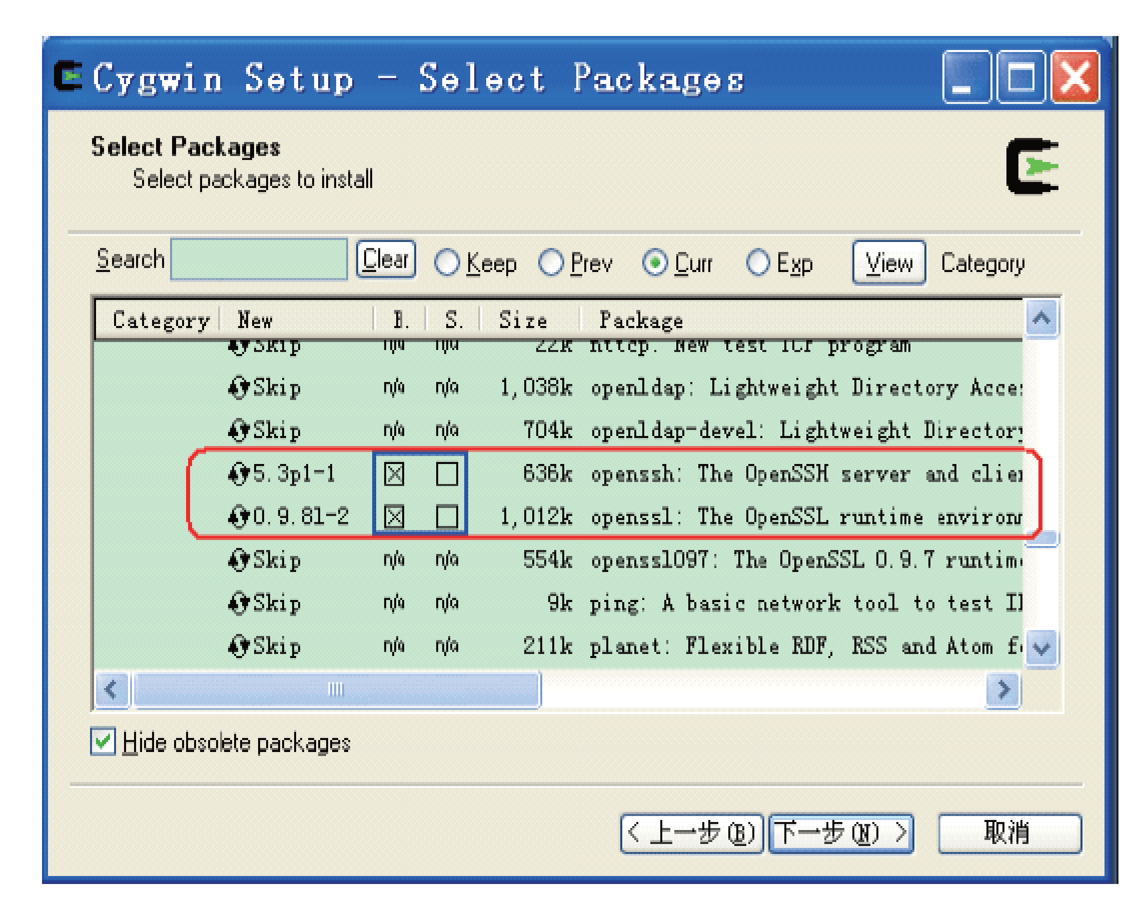


图4.15　保证“OpenSSL”被安装

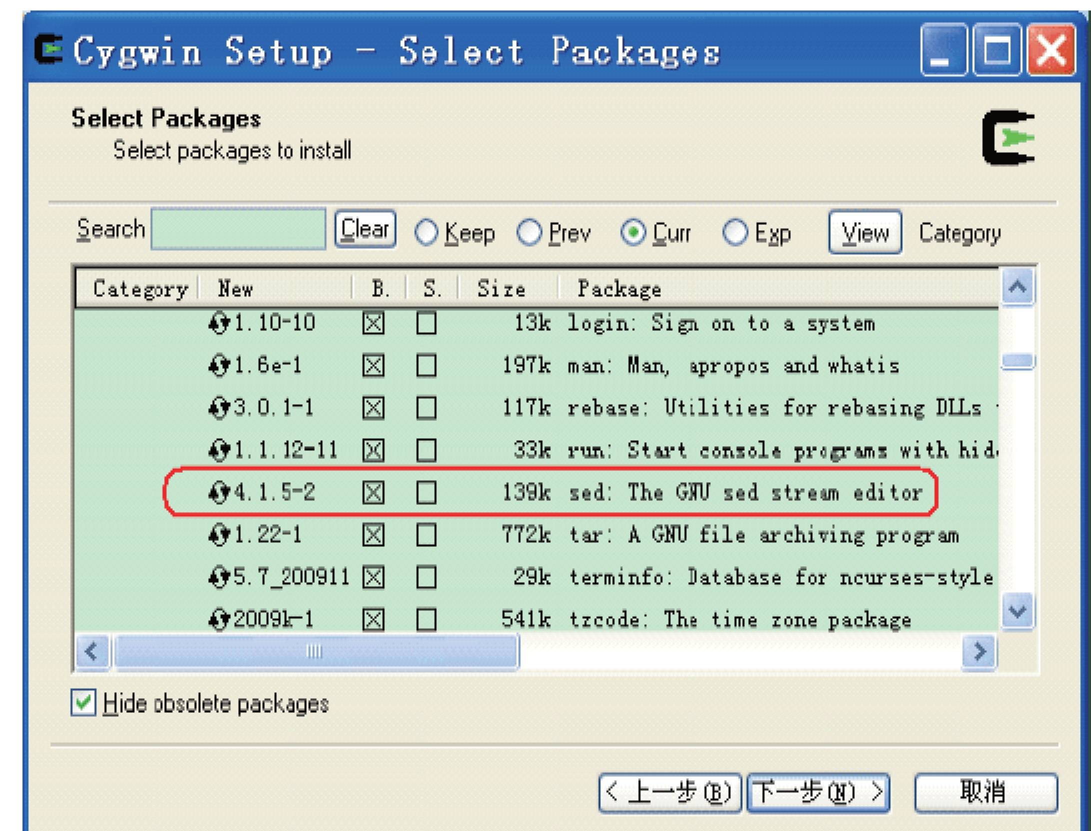


图4.16　安装“sed”

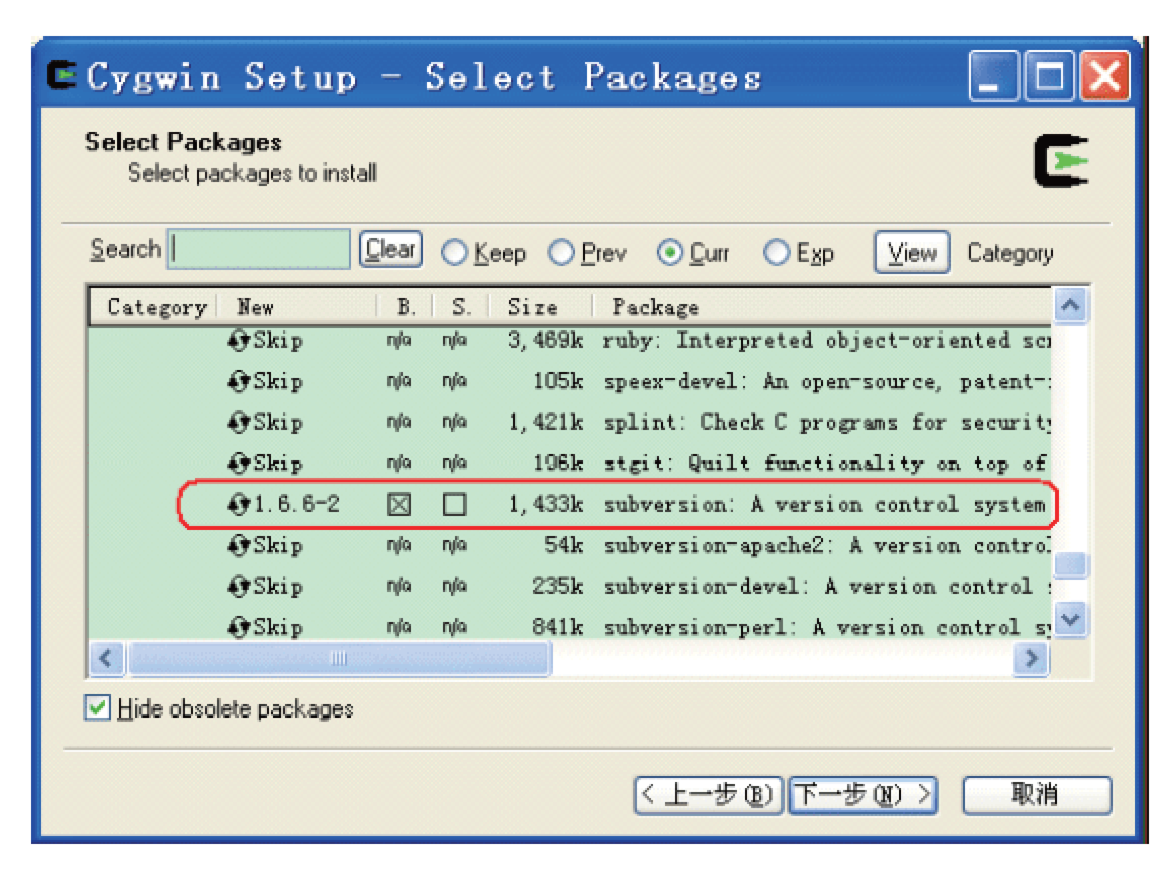


图4.17　建议安装“subversion”

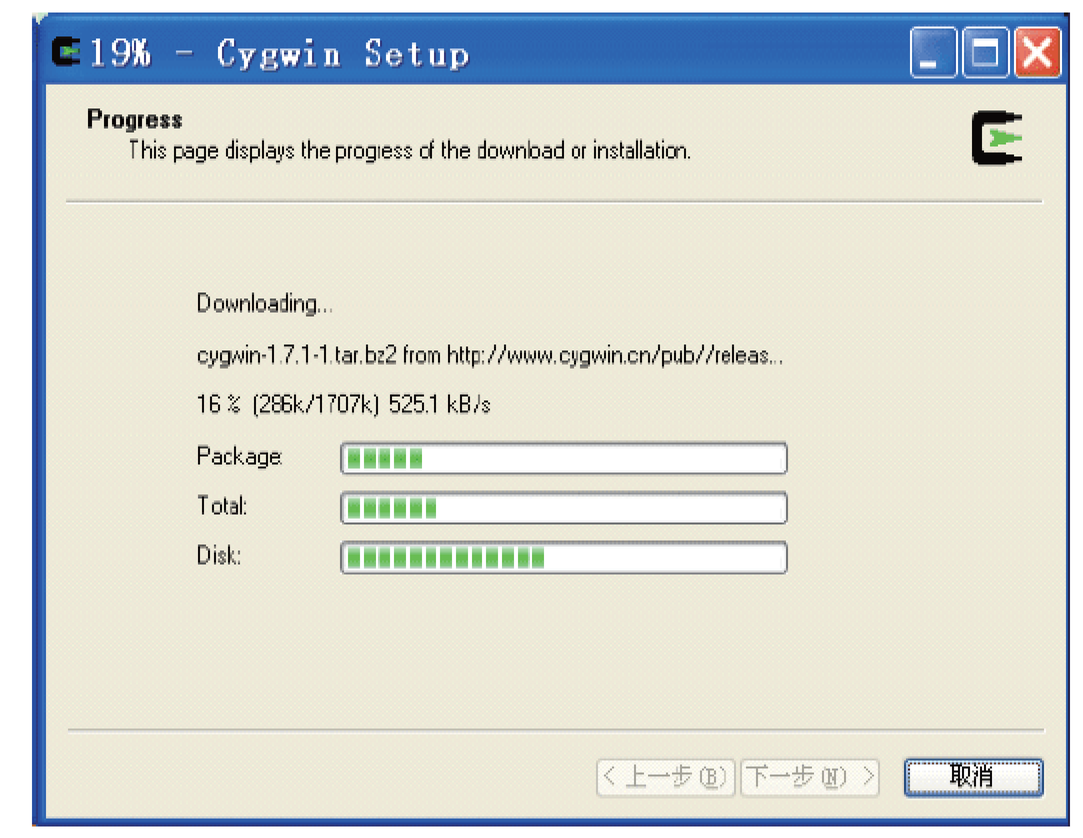


图4.18　下载安装包

当安装包下载完毕后，将会自动进入到如图4.19所示的对话框。

在图4.19所示的对话框中，选中“Create icon on Desktop”复选框，以方便直接从桌面上启动Cygwin，然后单击“完成”按钮。至此，Cygwin已经安装完，安装目录下的内容如图4.20所示。

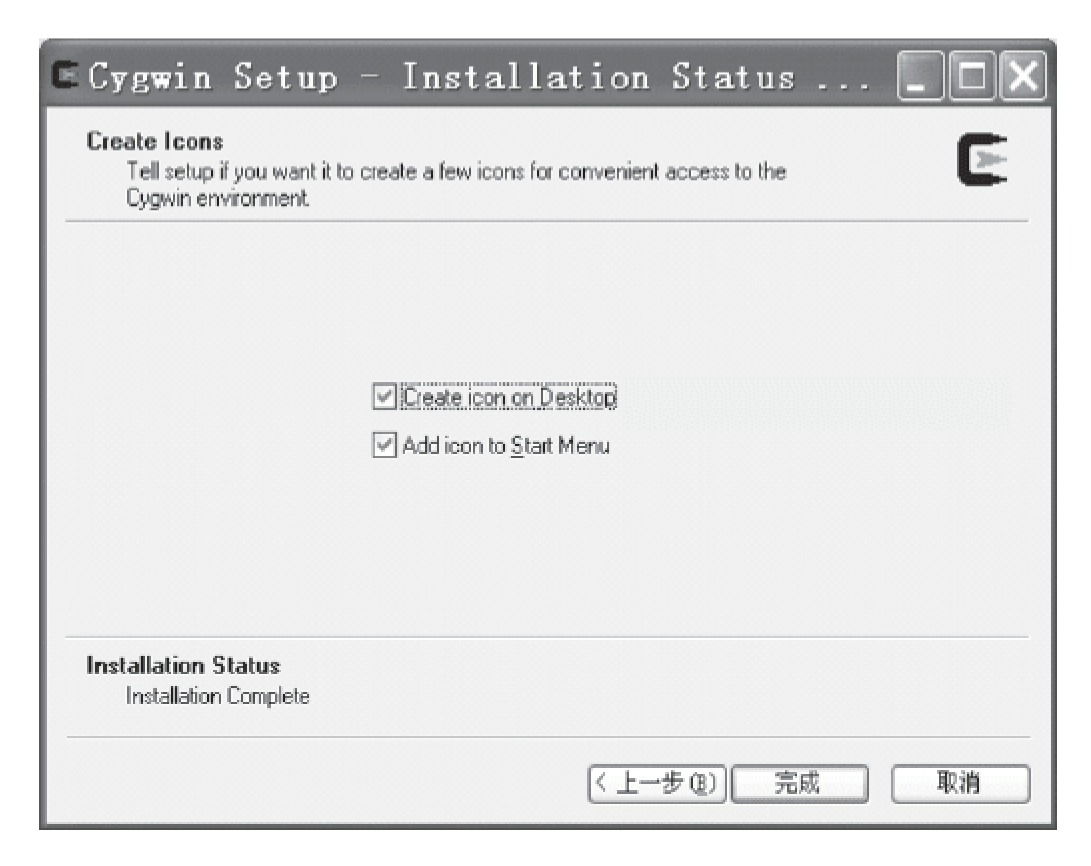


图4.19　在桌面上创建Cygwin图标

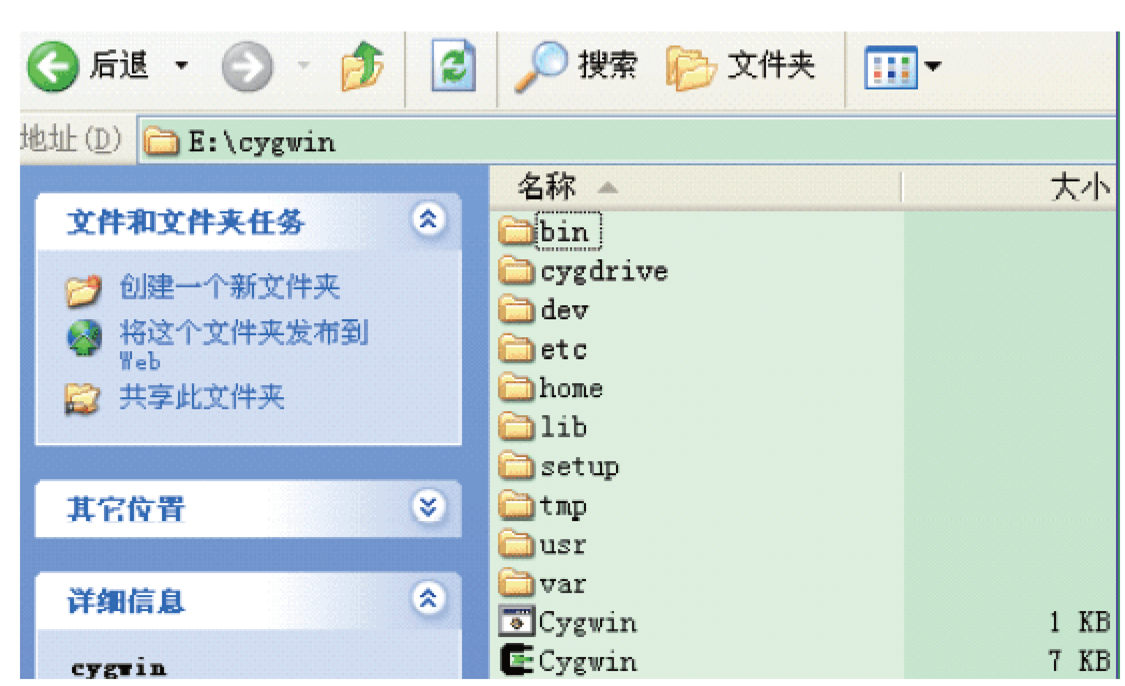


图4.20　Cygwin安装好后的目录

3. 配置环境变量

环境变量是在操作系统中一个具有特定名字的对象，它包含了一个或者多个应用程序将使用到的信息。用户通过设置环境变量来更好地运行进程。

需要配置的环境变量包括PATH 和JAVA_HOME。JAVA_HOME 指向JDK的安装目录；JDK的bin目录、Cygwin的bin目录以及Cygwin的usr\bin目录都必须添加到PATH环境变量中，如图4.21所示。

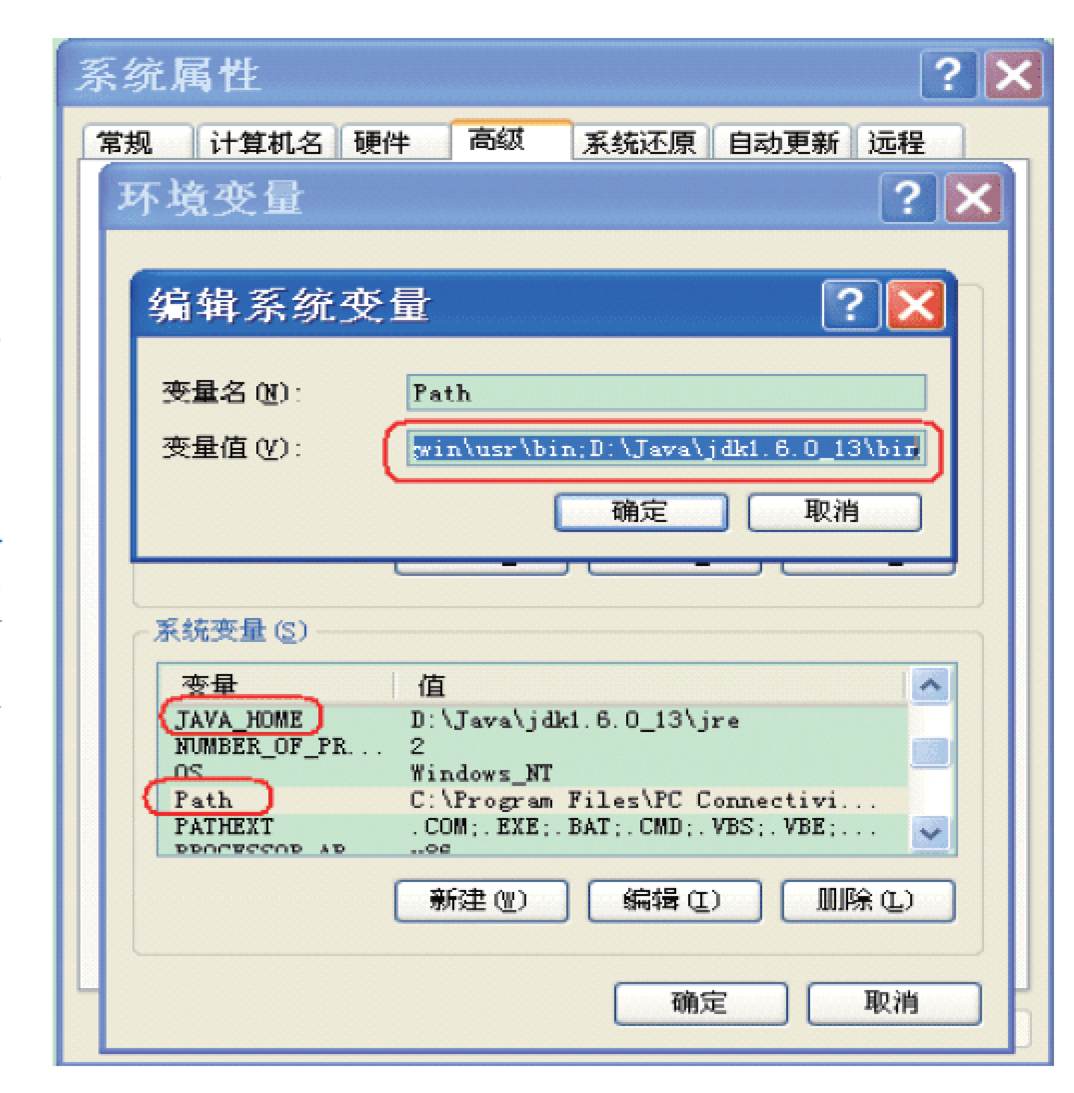


图4.21　在Windows上安装和配置Hadoop

4. 安装sshd服务

sshd是Cygwin OpenSSH壳程序，用于支持安全的SSH链接访问。双击桌面上的Cygwin图标，启动Cygwin，执行ssh-host-config命令，如图4.22所示。

在执行ssh-host-config 时，当要求输入yes/no时，选择输入no，如图4.23所示。

```
Copying skeleton files.
These files are for the user to personalise their cygwin experience.

They will never be overwritten nor automatically updated.

`./.bashrc' -> `/home/jian//.bashrc'
`./.bash_profile' -> `/home/jian//.bash_profile'
`./.inputrc' -> `/home/jian//.inputrc'
Your group is currently "passwd/group_GID_clash(gid/pgsid)". This
indicates that your gid is not in /etc/group, but the pgsid (primary
group associated with your SID) is in /etc/group.

The /etc/passwd (and possibly /etc/group) files should be rebuilt.
See the man pages for mkpasswd and mkgroup then, for example, run

mkpasswd -l [-d] > /etc/passwd
mkgroup  -l [-d] > /etc/group

Note that the -d switch is necessary for domain users.

jian@yijian ~
$ ssh-host-config
```

图4.22　执行ssh-host-config命令

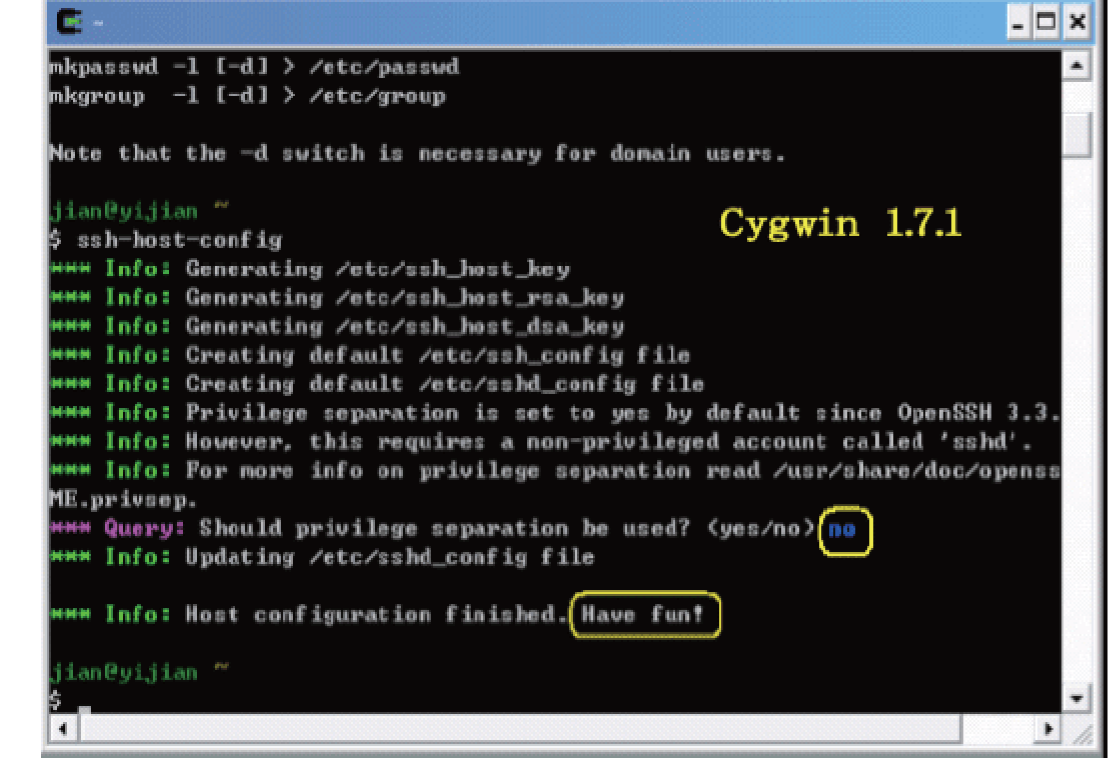


图4.23　在Windows上安装和配置Hadoop

若是Cygwin 1.7以前的版本，则ssh-host-config显示界面会如图4.24所示。

```
$ ssh-host-config
Generating /etc/ssh_host_key
Generating /etc/ssh_host_rsa_key
Generating /etc/ssh_host_dsa_key
Generating /etc/ssh_config file
Privilege separation is set to yes by default since OpenSSH 3.3.
However, this requires a non-privileged account called 'sshd'.
For more info on privilege separation read /usr/share/doc/openssh/R
.
Should privilege separation be used? (yes/no) no
Generating /etc/sshd_config file
Added ssh to C:\WINDOWS\system32\drivers\etc\services

Warning: The following functions require administrator privileges!

Do you want to install sshd as service?
(Say "no" if it's already installed as service) (yes/no) yes

Which value should the environment variable CYGWIN have when
sshd starts? It's recommended to set at least "ntsec" to be
able to change user context without password.
Default is "ntsec".  CYGWIN=ntsec

The service has been installed under LocalSystem account.
To start the service, call `net start sshd' or `cygrunsrv -S sshd'.

Host configuration finished. Have fun!
```

Cygwin 1.7之前的版本界面

图4.24　Cygwin 1.7以前版本

当看到出现“Have fun”时，一般表示sshd服务安装成功了，如图4.23所示。接下来，需要启动sshd服务。

5. 启动sshd服务

在桌面上的“我的电脑”图标上单击右键，单击“管理”菜单，进入Windows 计算机管理界面，如图4.25所示。

在图4.25所示的窗口中，右击“CYGWIN sshd”项，选择“启动”，启动CYGWIN sshd服务。启动成功后，如图4.26所示。

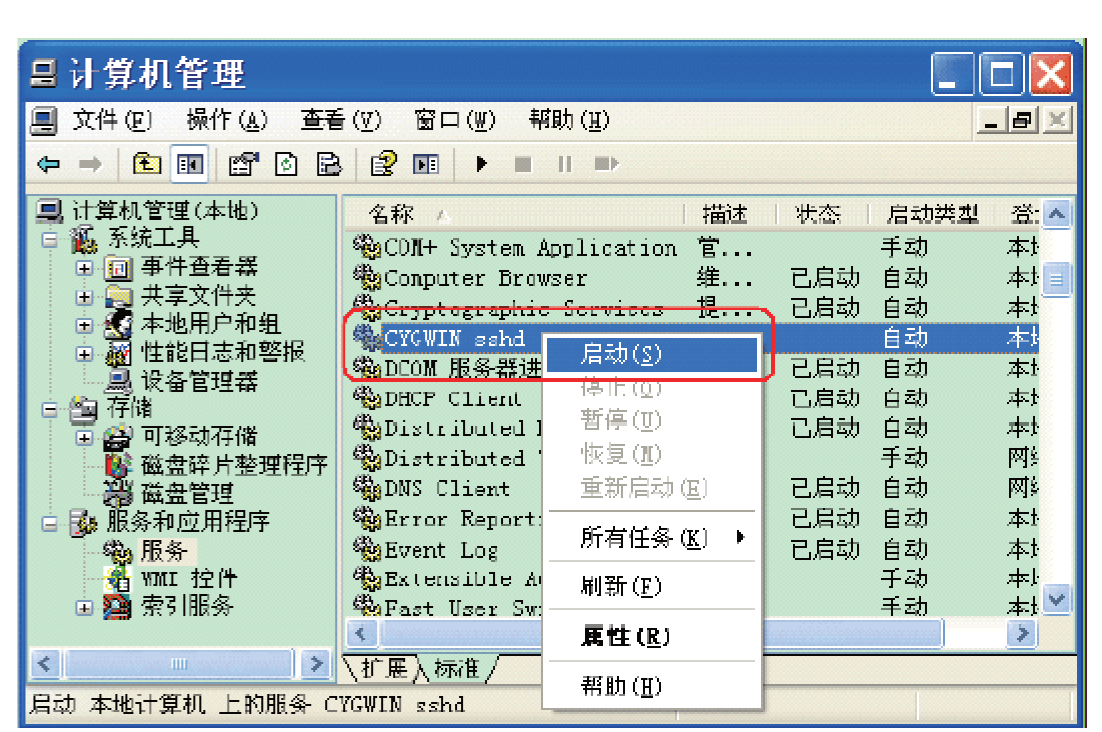


图4.25　Windows的计算机管理界面

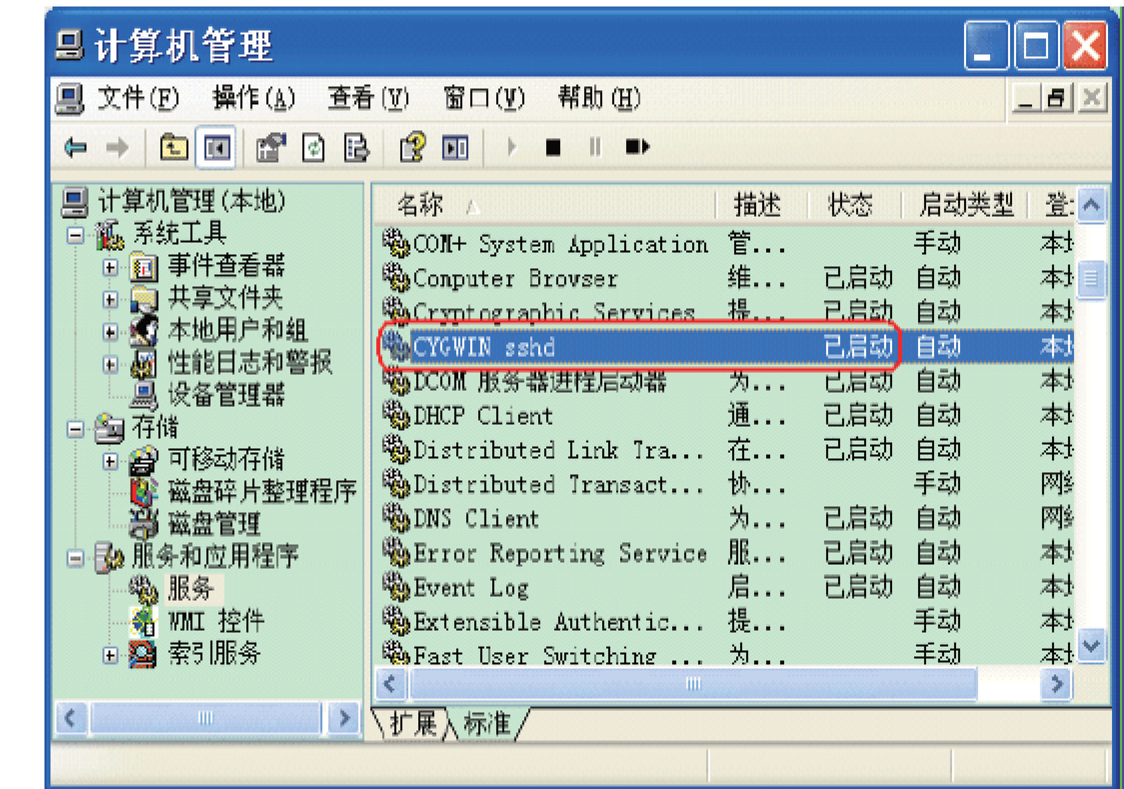


图4.26　启动“CYGWIN sshd”

当CYGWIN sshd的状态为“已启动”后，接下来就是配置ssh登录。

6. 配置ssh登录

执行ssh-keygen命令生成密钥文件，如图4.27所示。

在图4.27所示对话框中，需要输入时，直接按回车键即可。如果不出错，应当是需要三次按下回车键。接下来将生成authorized_keys文件，按图4.28所示操作即可。

```
~/.ssh
jian@yijian ~/.ssh
$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/jian/.ssh/id_rsa): 直接回车
Enter passphrase (empty for no passphrase): 这里也直接回车
Enter same passphrase again: 这里也直接回车
Your identification has been saved in /home/jian/.ssh/id_rsa.
Your public key has been saved in /home/jian/.ssh/id_rsa.pub.
The key fingerprint is:
e1:a3:9d:4e:b5:3b:95:f0:66:1c:1c:10:58:ca:41:e2 jian@yijian
The key's randomart image is:
+--[ RSA 2048]----+
```

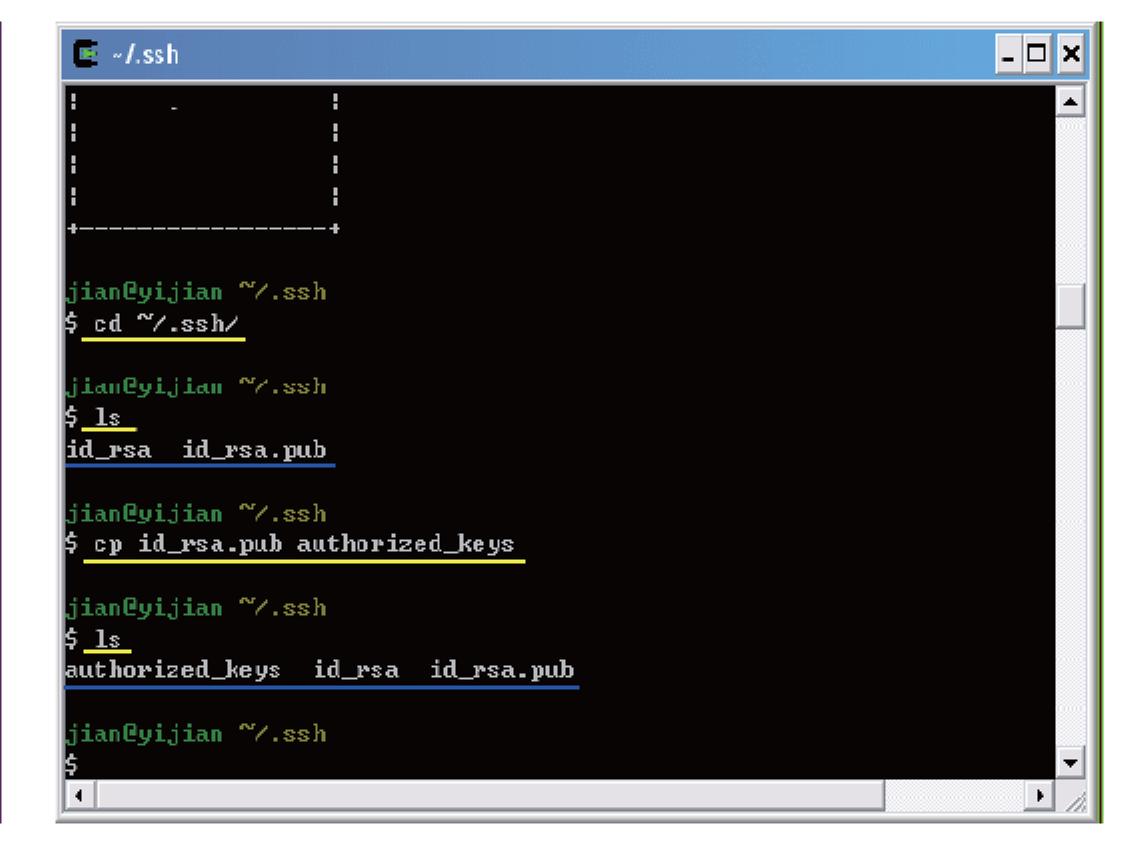


图4.27　生成密钥文件　　　　图4.28　生成authorized_keys文件

至此，只需执行以下两步操作，即可生成authorized_keys文件。

```
cd ~/..ssh/
cp id_rsa.pub authorized_keys
```

完成上述操作后，执行exit命令先退出Cygwin窗口。如果不执行这一步操作，则下面的操作可能会遇到错误。

接下来，重新运行Cygwin，执行ssh localhost命令。在第一次执行ssh localhost时，会有如图4.29所示的提示，输入yes，然后按回车键即可。

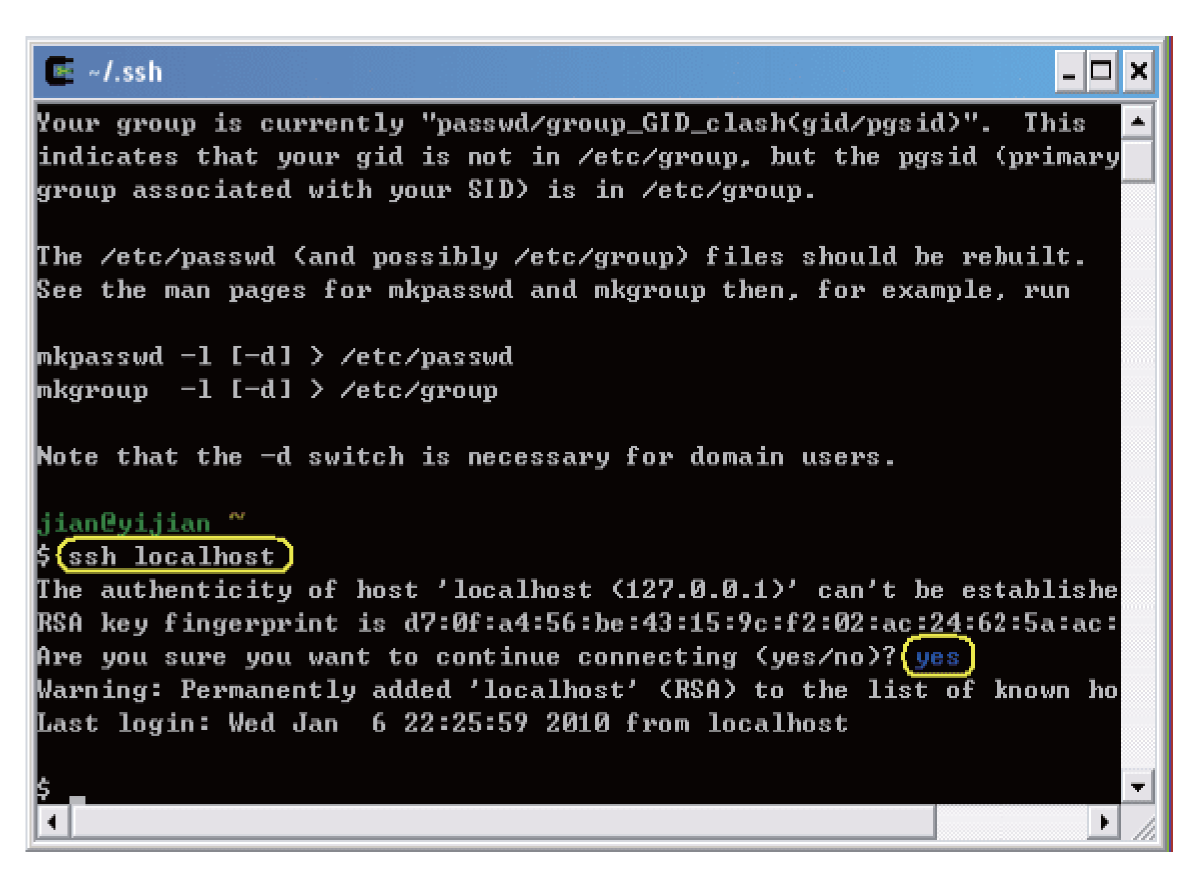


图4.29　执行ssh localhost命令

在这步操作中，如果是Windows域用户，则可能会遇到问题，错误信息如图4.30所示。

```
$ ssh localhost
Last login: Thu Jan  7 10:23:09 2010 from localhost
      3 [main] -bash 5016 E:\cygwin\bin\bash.exe: *** fatal e
amically determine load address for 'WSAGetLastError' (handle
 error 126
Connection to localhost closed.
```

图4.30　错误提示信息

该错误暂时没有解决办法，可关注Hadoop技术论坛中的贴：http://bbs.hadoopor.com/thread-348-1-1.html（Cygwin 1.7.1版本ssh问题），跟进问题的解决情况。如果操作成功，执行who命令时，就会看到如图4.31所示的信息。

图4.31　执行who命令

至此，配置ssh登录成功，接下来就可以进行Hadoop的安装。

7. 下载Hadoop安装包

Hadoop 安装包下载地址是http://labs.xiaonei.com/apache-mirror/hadoop/core/ha doop-0. 20.1/hadoop-0.20.1.tar.gz，下载安装包。

8. 安装Hadoop

将下载Hadoop安装包hadoop-0.20.1.tar.gz解压到D:\hadoop\run目录（可以修改成其他目录）下，如图4.32所示。

修改Hadoop的配置文件。配置文件位于conf 子目录下，包括hadoop-env.sh、core-site.xml、hdfs-site.xml和mapred-site.xml四个文件。在Cygwin环境下，masters和slaves两个文件不需

要修改。

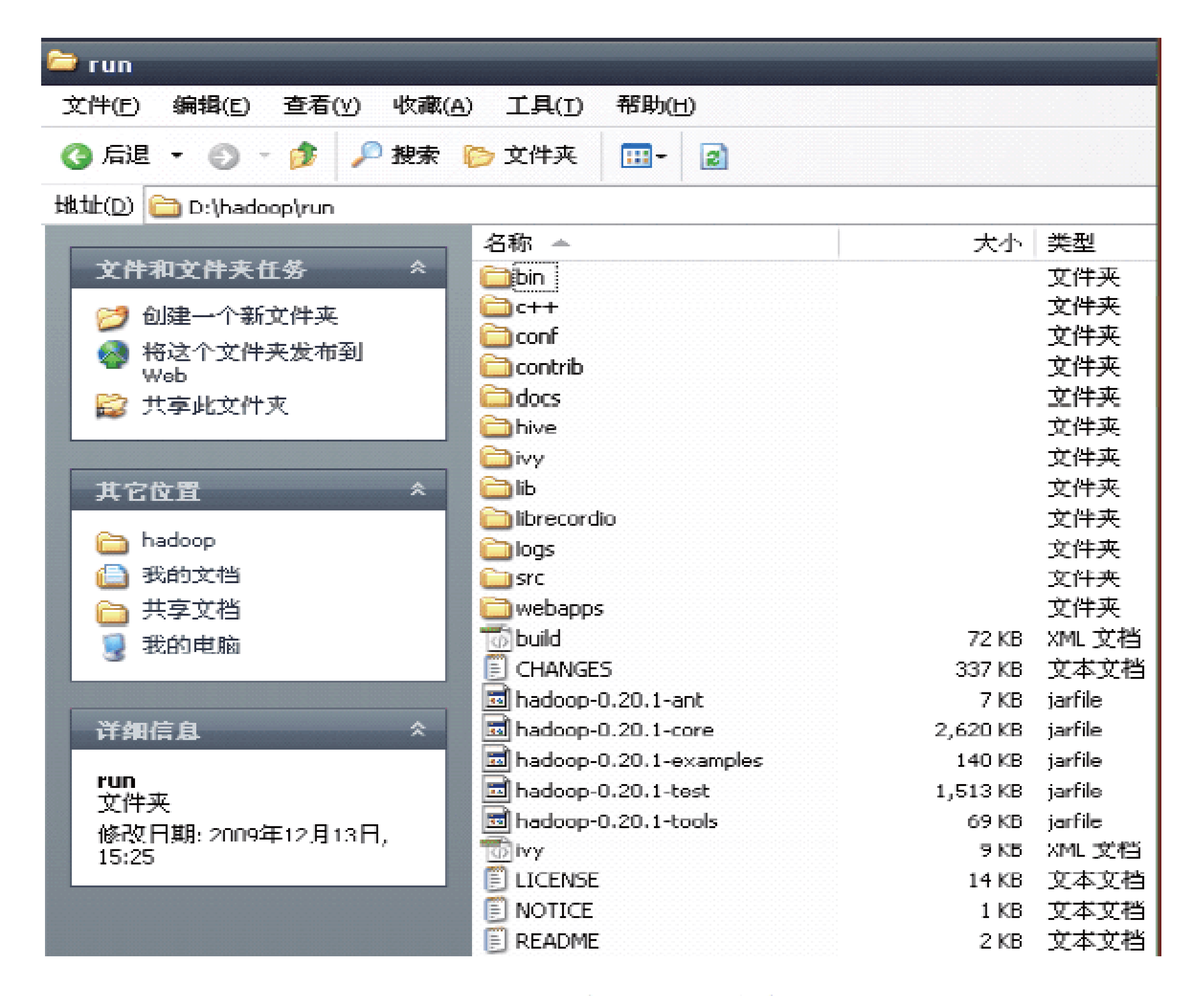


图4.32 解压安装包

（1）修改hadoop-env.sh

将JAVA_HOME 修改成JDK的安装目录即可。注意，JDK必须是1.6 或以上版本。

（2）修改core-site.xml

为简化core-site.xml配置，将D:\hadoop\run\src\core目录下的core-default.xml文件复制到D:\hadoop\run\conf目录下，并将core-default.xml文件名改成core-site.xml。修改fs.default.name的值，如图4.33所示。

```
<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:8888</value>
  <description>The name of the default file system
  scheme and authority determine the FileSystem im
  uri's scheme determines the config property (fs.
  the FileSystem implementation class.  The uri's
  determine the host, port, etc. for a filesystem.
</property>
```

图4.33 修改fs.default.name的值

图4.33中的端口号为8888，可以改成其他未被占用的端口。

（3）修改hdfs-site.xml

为简化hdfs-site.xml配置，将D:\hadoop\run\src\hdfs目录下的hdfs-default.xml文件复制到D:\hadoop\run\conf目录下，并将hdfs-default.xml文件名改成hdfs-site.xml。不需要再做其他修改。

（4）修改mapred-site.xml

为简化mapred-site.xml 配置，将D:\hadoop\run\src\mapred目录下的mapred-default.xml文件复制到D:\hadoop\run\conf目录下，并将mapred-default.xml文件名改成mapred-site.xml。

图4.34中的端口号为9999，可以改成其他未被占用的端口。至此，表示Hadoop安装完毕，可以开始体验Hadoop了。

```
<property>
  <name>mapred.job.tracker</name>
  <value>localhost:9999</value>
  <description>The host and port that the MapReduce job
  at.  If "local", then jobs are run in-process as a si
  and reduce task.
  </description>
</property>
```

图4.34　修改端口号

9. 启动Hadoop

在Cygwin中，进入Hadoop的bin目录，运行/start-all.sh来启动Hadoop。在启动成功之后，可以执行/hadoop fs –ls/命令，查看Hadoop的根目录，如图4.35所示。

```
/cygdrive/d/hadoop/run/bin
ondarynamenode-yijian.out' for reading: No such file or directory
starting jobtracker, logging to /cygdrive/d/hadoop/run/bin/../logs/hadoo
obtracker-yijian.out
localhost: starting tasktracker, logging to /cygdrive/d/hadoop/run/bin/.
adoop-jian-tasktracker-yijian.out
localhost: /cygdrive/d/hadoop/run/bin/hadoop-daemon.sh: line 117: /cygdr
doop/run/bin/../logs/hadoop-jian-tasktracker-yijian.out: Permission deni
localhost: head: cannot open `/cygdrive/d/hadoop/run/bin/../logs/hadoop-
ktracker-yijian.out' for reading: No such file or directory
jian@yijian /cygdrive/d/hadoop/run/bin
$ jps
5332 JobTracker
4192 Jps
4220 NameNode

jian@yijian /cygdrive/d/hadoop/run/bin
$ ./hadoop fs -ls /
Found 4 items
drwxr-xr-x   - jian supergroup          0 2009-12-18 19:28 /tmp
$
```

图4.35　启动Hadoop后的界面

4.3.2　在Linux上安装与配置Hadoop

在Linux操作系统上安装与配置Hadoop的步骤与在Windows下类似。以下操作均在虚拟机VMWare 10.0上安装的Ubuntu 12.04（Linux的一个版本）环境下实现。首先应该在Ubuntu下创建Hadoop用户。

先在Ubuntu的控制台上查看系统版本以及操作系统位数，如图4.36所示。

```
hadoop@ubuntu-vm-server:~$ uname -a
Linux ubuntu-vm-server 3.2.0-57-generic-pae #87-Ubuntu SMP Tue Nov 12 21:57:43 U
TC 2013 i686 i686 i386 GNU/Linux
hadoop@ubuntu-vm-server:~$
```

图4.36　Ubuntu的控制台

如果显示X86_64说明是64位内核，显示i386、i686说明是32位内核。

1. 安装JDK

实验使用的JDK文件为jdk–7u45–linux–i586.bin，下载完毕后放在Home目录下。

（1）进入存放JDK文件的目录并解压

执行dk–7u45–linux–i586.gz $tar －zxvf *.gz命令。

（2）创建/usr/lib/jvm文件

执行sudo mkdir /usr/lib/jvm命令，之后可看到如图4.37所示的界面。

```
adminator@adminator-virtual-machine:/$ cd /usr/lib
adminator@adminator-virtual-machine:/usr/lib$ sudo mkdir jvm
[sudo] password for adminator:
adminator@adminator-virtual-machine:/usr/lib$
```

图4.37　创建/usr/lib/jvm文件

复制解压后的jdk1.7.0_45到/usr/lib/jvm目录。

执行sudo cp －r jdk1.7.0_45 /usr/lib/jvm命令后可看到如图4.38所示界面。

```
adminator@adminator-virtual-machine:/$ cd home
adminator@adminator-virtual-machine:/home$ ls
adminator
adminator@adminator-virtual-machine:/home$ cd
adminator@adminator-virtual-machine:~$ ls
Desktop     examples.desktop             Music      Templates
Documents   jdk1.7.0_45                  Pictures   Videos
Downloads   jdk-7u45-linux-i586.tar.gz   Public
adminator@adminator-virtual-machine:~$ sudo cp -r jdk1.7.0_45 /usr/lib/jvm
adminator@adminator-virtual-machine:~$
```

图4.38　执行sudo cp–r jdk 1.7.0–45/usr/lib/jvm命令

（3）赋予文件读、写和执行的权限

执行sudo chmod 777 /usr/lib/jvm命令。

（4）修改环境变量（对计算机的所有用户生效）

执行sudo gedit /etc/profile命令。

（5）使用source更新/etc/profile

执行source /etc/profile命令。

使用java －version命令查看Java版本号，如图4.39所示。

```
adminator@adminator-virtual-machine:~$ java -version
java version "1.7.0_45"
Java(TM) SE Runtime Environment (build 1.7.0_45-b18)
Java HotSpot(TM) Client VM (build 24.45-b08, mixed mode)
```

图4.39　使用java–version命令查看Java版本号

2. 创建Hadoop用户组和用户

（1）创建Hadoop用户组

执行sudo addgroup hadoop命令。

（2）创建Hadoop用户并设置密码

执行sudo adduser －ingroup hadoop hadoop命令。

（3）赋予用户权限

执行 sudo gedit /etc/sudoers命令。打开/etc/sudoers文件，给Hadoop用户赋予与root用户同样的权限。在root ALL=(ALL:ALL) ALL下添加hadoop ALL=(ALL:ALL) ALL语句，如图4.40所示。

```
# User privilege specification
root     ALL=(ALL:ALL) ALL
hadoop   ALL=(ALL:ALL) ALL
```

图4.40　赋予用户权限

3. 切换Hadoop用户，安装并配置ssh

使用Hadoop用户配置无密码访问本机。

（1）安装ssh

SSH是Secure Shell的缩写，是建立在应用层和传输层上的安全协议。因为Hadoop运行过程中需要管理远程的Hadoop守护进程，如果在安装Linux虚拟机时没有安装ssh，可以使用下面的命令来安装。

sudo apt-get install openssh-server

在安装过程中可能会遇到一些问题，或需要安装其他的程序，这时可到网上寻找帮助或下载相应的安装程序。比如需要安装Libck-connector0_0.4.5-3ubuntu0.1_i386.deb，则可以先下载它然后安装就行了。

（2）配置ssh免密码登录

ssh生成密钥方式有RSA（专用密钥）和DSA（公用密钥）两种，默认情况下采用RSA方式。

① 生成密钥对，执行如下命令。

```
ssh-keygen -t rsa
```

在询问保存路径时，直接按回车键采用默认路径。提示要为生成的秘钥输入passphrase时，直接按回车键将其设定为空密码。连续按三次回车键，之后会在~/.ssh/目录下生成两个文件：id_rsa（私用密钥）和id_rsa.pub（公用密钥），这两个文件是成对出现的，可用ls查看。

② 进入~/.ssh/目录下，将id_rsa.pub追加到authorized_keys授权文件中。开始是没有authorizaed_keys文件的。

```
cd ~/.ssh
ls
```

③ 将id_rsa.pub追加到authorized_keys授权文件中，命令如下。

```
cat id_rsa.pub >> authorized_keys
```

或

```
 $cp id_rsa.pub authorized_keys
```

完成之后就可以无密码登录本机了。

④ 登录localhost，命令如下。

```
ssh localhost
```

输入yes，如sshd没有启动，则会显示22号端口打不开，此时找出错误，重新配置即可。当进行伪分布模式的配置时，配置完成之后要执行exit退出命令，因为此时控制的是远程的机器，需要执行退出命令才能重新控制本机。

4. Hadoop的安装与部署

将下载的Hadoop-2.2.0软件下载到本机，并解压到home目录下。

（1）命令行解压

Linux解压tar的命令如下。

```
$tar -zxvf *.tar.gz
```

参数含义如下。

- –c：打包
- –t：查看tar目录里的内容
- –j：用bzip2压缩
- –p：保留文件原属性
- –P：使用绝对属性
- –N：压缩指定日期（yyy/mm/dd）以后用的文件
- –x：解开压缩文件
- –z：用gzip压缩
- –v：压缩过程同时显示文件列表
- –f：文件名。该参数必须是最后一个参数，后面接文件名

改名操作，步骤如下。

执行sudo mv hadoop–2.2.0 hadoop命令。

将Hadoop目录的属主用户设为hadoop。

执行sudo chown – R hadoop:hadoop hadoop命令。

Hadoop有三种运行模式：单机模式、单机伪分布模式式与完全分布模式。虽然第一、二种运行模式无法充分展现云计算的优势，也没有实际的应用意义，但是它们在程序的测试与调试过程中还是发挥作用的。

（2）单机配置方式

单机模式不用配置，选择这种方式，Hadoop可以作为一个单独的Java进程，常被用于调试程序。

（3）单机伪分布式模式配置方式

这种模式可以被当作是只有一个节点的集群，这个节点既充当Master，也充当Slave；既可以看作NameNode，也可以看作DataNode；既能作为JobTracker，也能作为TaskTracker。

在这种模式下修改4个配置文件即可。这四个配置文件均位于/etc/hadoop目录下。

① 配置hadoop–env.sh

设置JAVA_HOME变量为：

```
export JAVA_HOME=/usr/lib/jvm/jdk1.7.0_45
```

为了防止出错，AVA_HOME变量的查看可通过下面的命令实现。执行命令后界面如图4.41所示。

```
echo $JAVA_HOME
```

```
hadoop@adminator-virtual-machine:~$ echo $JAVA_HOME
/usr/lib/jvm/jdk1.7.0_45
```

图4.41　在Linux上安装与配置Hadoop

② 配置core–site.xml

在configuration中添加如下语句。

```
<property>
```

```
<name>hadoop.tmp.dir</name>
<value>/home/hadoop/hadoop-2.2.0/hadoop-datastore/tmp</value>
<description>A base for other temporarydirectories.</description>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:8010</value>
<description>The name of the default file system.  A URI whose
scheme and authority determine the FileSystem implementation. The
uri's scheme determines the config property (fs.SCHEME.impl) naming
the FileSystem implementation class. The uri's authority is used to
determine the host, port, etc. for a filesystem.</description>
</property>
```

③ 配置hdfs-site.xml

在configuration添加如下语句。

```
<property>
<name>dfs.replication</name>
<value>1</value>
<description>Default block replication.
The actual number of replications can be specified when the file iscreated.
The default is used if replication is not specified in create time.
</description>
</property>
```

④ 配置mapred-site.xml

对于hadoop2.0以前的版本，配置mapred-site.xml文件；对于hadoop2.0版本，配置mapred-site.xml.template文件。

在configuration之间添加如下语句。

```
<property>
<name>mapred.job.tracker</name>
<value>localhost:54311</value>
<description>The host and port that the MapReduce job tracker runs
at. If "local", thenjobs are run in-process as a single map
and reduce task.
</description>
</property>
<property>
<name>mapred.map.tasks</name>
```

```
<value>10</value>
<description>As a rule of thumb, use 10x the number of slaves(i.e., number
of tasktrackers).
</description>
</property>
<property>
<name>mapred.reduce.tasks</name>
<value>2</value>
<description>As a rule of thumb, use 2x the number of slaveprocessors (i.e.,
number of tasktrackers).
</description>
</property>
```

⑤ 格式化NameNode

启动Hadoop前，需要先格式化Hadoop的文件系统HDFS，执行如下命令。

```
bin/hdfs -format
```

⑥ 启动Hadoop

如果格式化成功，则可以启动Hadoop守护进程：输入命令

```
bin/start-dfs.sh
```

或

```
bin/start-yarn.sh
```

启动进程。

相应地，停止Hadoop守护进程的命令是

```
bin/stop-dfs.sh
```

和

```
bin/stop--yarn.sh
```

⑦验证测试

验证需要在启动模式下进行。

首先输入jps，可以查看端口号，如图4.42所示。

```
hadoop@adminator-virtual-machine:~/hadoop-2.2.0$ jps
3746 Jps
3676 NodeManager
3554 ResourceManager
3206 DataNode
3376 SecondaryNameNode
3092 NameNode
```

图4.42　查看端口号

然后打开浏览器，输入下列网址。如果能够正常浏览，说明安装成功。

http://localhost:8088

http://localhost:50070，

第1个网址可链接到Hadoop资源管理器，第2个网址可链接到NameNode的Web页面，对HDFS进行检测。在Hadoop上查看端口情况，如图4.43~图4.45所示。

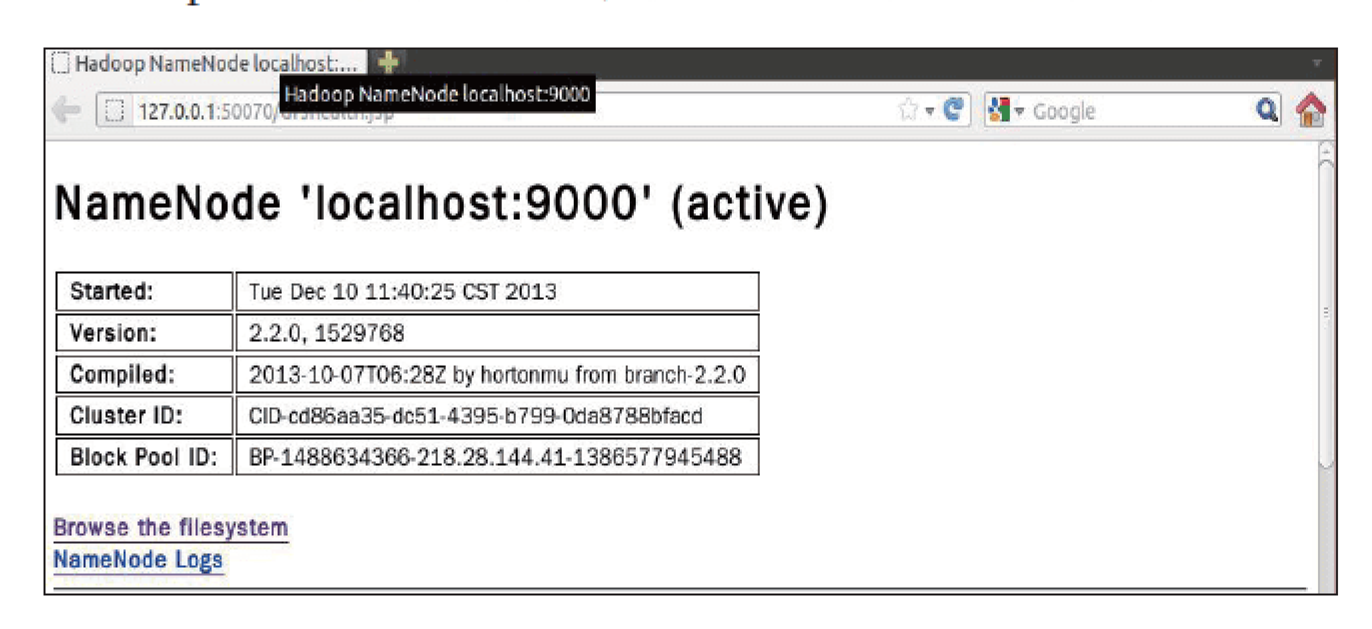


图4.43　端口情况

```
user@localhost:~
文件(F)  编辑(E)  查看(V)  搜索(S)  终端(T)  帮助(H)
[user@localhost ~]$ lsof -i:9000
COMMAND  PID USER   FD   TYPE DEVICE SIZE/OFF NODE NAME
java    2077 user   66u  IPv6  22406      0t0  TCP master:cslistener (LISTEN)
java    2077 user   81u  IPv6  46526      0t0  TCP master:cslistener->master:49146 (ESTABLISHED)
java    2325 user   78u  IPv6  47412      0t0  TCP master:49146->master:cslistener (ESTABLISHED)
[user@localhost ~]$
```

图4.44　在Linux上安装与配置

```
user@localhost:~
文件(F)  编辑(E)  查看(V)  搜索(S)  终端(T)  帮助(H)
[user@localhost ~]$ netstat -nap | grep 9000
(Not all processes could be identified, non-owned process info
 will not be shown, you would have to be root to see it all.)
tcp        0      0 ::ffff:192.168.162.13:9000  :::*                        LISTEN      2077/java
tcp        0      0 ::ffff:192.168.162.13:49178 ::ffff:192.168.162.13:9000  ESTABLISHED 2325/java
tcp        0      0 ::ffff:192.168.162.13:9000  ::ffff:192.168.162.13:49178 ESTABLISHED 2077/java
tcp        0      0 ::ffff:192.168.162.13:49176 ::ffff:192.168.162.13:9000  TIME_WAIT   -
tcp        0      0 ::ffff:192.168.162.13:49177 ::ffff:192.168.162.13:9000  TIME_WAIT   -
tcp        0      0 ::ffff:192.168.162.13:49174 ::ffff:192.168.162.13:9000  TIME_WAIT   -
[user@localhost ~]$
```

图4.45　在Linux上安装与配置

4.4　Hadoop应用案例

本节以Last.fm和Facebook为例说明Hadoop在这些软件上的应用。

4.4.1　Last・fm

Last・fm是世界上最大的社交音乐平台。创建于2002年，提供网络电台和网络音乐服务。Last・fm音乐库里有约超过1000万个歌手和超过1亿首歌曲。每个月，全世界250个国家，超过2000万人在这里寻找、收听、谈论自己喜欢的音乐。这些数字不断增长，产生大量数据。2006年初，Last・fm开始使用Hadoop，几个月后投入实际应用。Hadoop是Last・fm基础平台的关键组件，有2个Hadoop集群、50台计算机、300个内核、100TB的硬盘空间。在集群上，运行数百种日常作业，包括日志文件分析、A/B测试评测、即时处理和图表生成。图4.46所示为Last・fm的图标。

图4.46　Llast.fm图标

Last・fm创建于2002年，它是一个提供网络电台和网络音乐服务的社交网络。每个月有2500万人使用Last・fm，会产生大量数据。

2006年初，Last・fm开始使用Hadoop，几个月后投入实际应用。Hadoop是Last・fm基础平台的关键组件，有2个Hadoop集群，50台计算机，300个内核，100TB的硬盘空间。在集群上，运行数百种日常作业，包括日志文件分析，A/B测试评测，即时处理和图表生成。

1. 图表生成

图表生成是Hadoop在Last・fm的第一个应用，如图4.47所示。

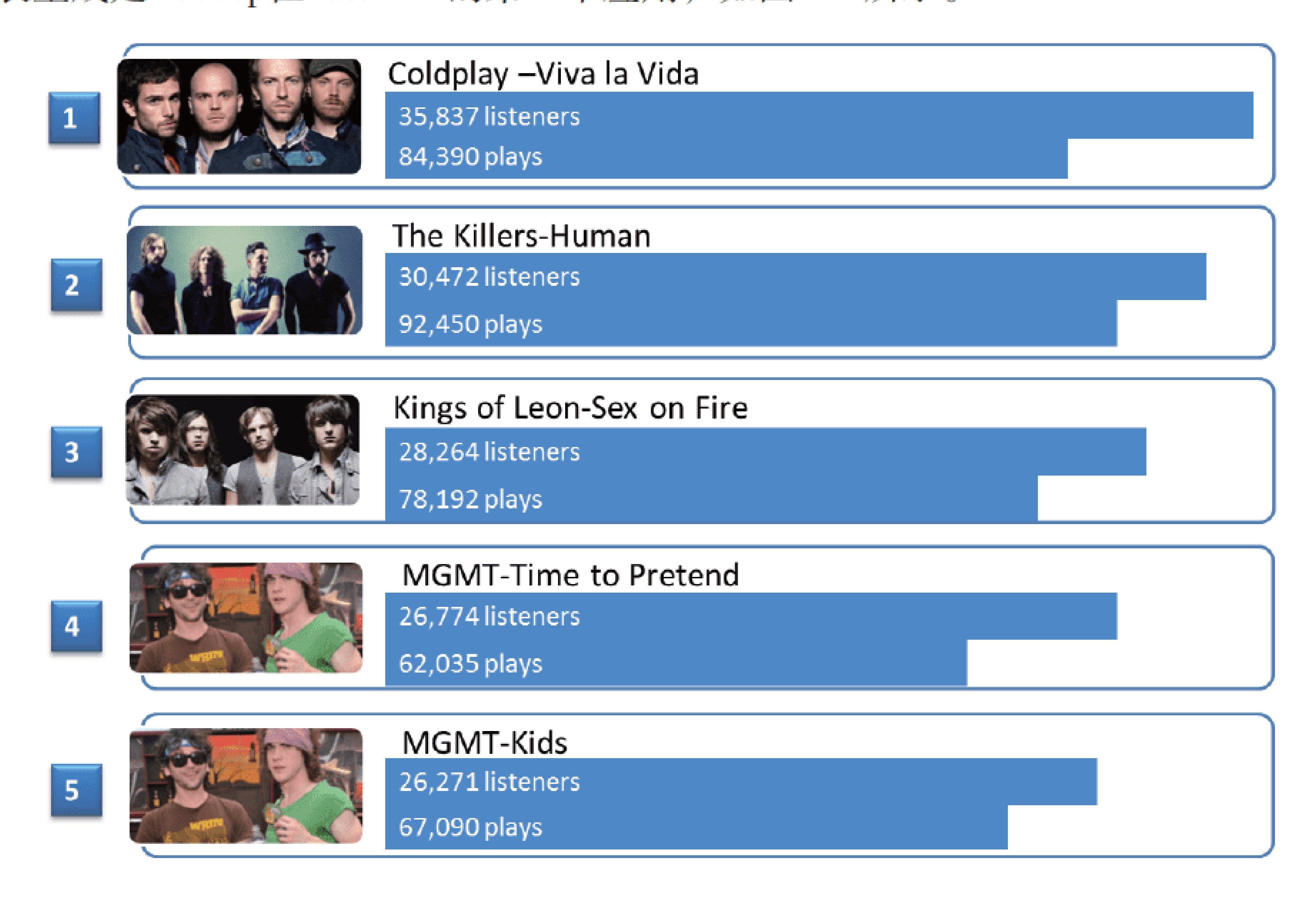


图4.47　Last・fm音乐排行统计图表

2. 数据从哪里来

Last・fm有两种收听数据：用户播放自己的音乐，如使用PC或者其他设备来播放mp3，这种信息通过Last・fm的客户端或者第三方应用发送到Last・fm，这一类叫scrobble收藏数据；用户收听Last・fm网络电台的节目，以及在听节目时候的喜好，跳过，禁止等操作信息，这一类叫 radiolisten电台收听数据。

3. 数据存储

收听数据被发送到Last・fm，经过验证和转换，形成一系列由空格分隔的文本文件，包含用户ID（userid），音乐ID（trackid），这首音乐被收藏的次数（scrobble），这首音乐在电台中收听的次数（radio），被跳过的次数（skip）。真实数据达到GB级别，有更多属性字段。

4. 数据处理

UniqueListeners作业：统计收听某一首歌的不同用户数。也就是说，有多少个用户听过某个歌。如果用户重复收听，只算一次。

Sum作业：每首歌的收听总数、收藏总数、电台收听总数、被跳过的总数。

合作作业：每首歌被多少不同用户收听总数、收听总数、收藏总数、电台收听总数、被跳过的总数。

这些数据会被用来制作周排行榜等，然后在Last.fm主站上显示出来。

5. 总结

Hadoop已经成为Last・fm基础框架的一个重要部件，它用于产生和处理各种各样的数据集，如网页日志信息和用户收听数据。为了让读者能够掌握主要的概念，这里讲述的例子已经被大大地简化，而在实际应用中，输入的数据具有更复杂的结构，并且数据处理的代码也更加繁琐。虽然Hadoop本身已经足够成熟，可以支持实际应用，但它仍在被开发人员积极地开发，并且Hadoop社区每周都会为它增加新的特性及提升它的性能。Last・fm很高兴是这个社区的一份子，是代码和新想法的贡献者，同时也是对大量开源技术进行利用的终端用户。

4.4.2 Facebook

图4.48所示为Facebook的图标。

图4.48 Facebook图标

1. 背景

Facebook是世界著名的大型社交网站，有3亿以上的用户活跃在此网站上，其中有10%左右的用户每天至少更新一次自己的状态；每个月用户累计上传超过10亿张图片和一千万个视频；每周累计共享10亿条内容，其中包括个人日志、网页链接、热点新闻、热门微博等。如此巨大的数据量都是Facebook需要存储和处理的，而且每天还要新增加4TB压缩后的数据，需要扫描高达135TB的数据，在集群上执行Hive任务超过7500次，每小时需要进行8万次计算……所以高性能的云平台对Facebook而言至关重要，而Facebook采用Hadoop平台，主要负责完成日志处理、推荐系统和数据仓库等各方面的工作。

2. 数据存储现状

Facebook将海量数据存储在数据仓库上，这个数据仓库是利用Hadoop/Hive搭建的，拥有4800个内核，5.5PB的存储量，单节点存储量达12TB，其两层网络拓扑如图4.49所示。Facebook中的MapReduce集群可以动态变化，它基于负载情况和集群节点之间的配置信息可动

态移动。

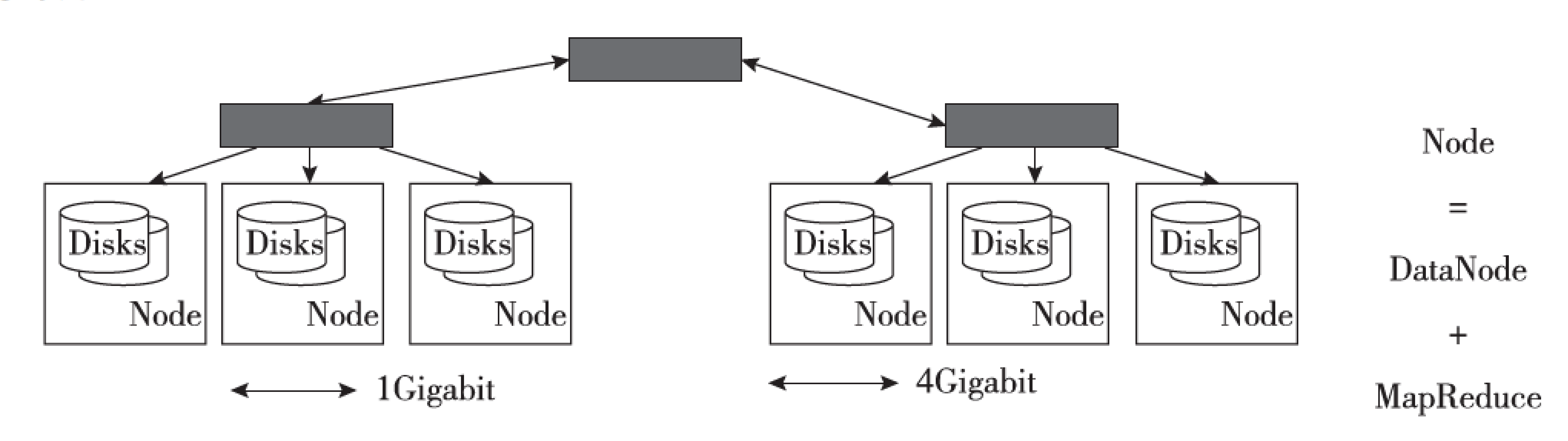


图4.49　集群的网络拓扑图

3. 数据仓库架构

图4.50为Facebook的数据仓库架构，在这个架构中，网络服务器和内部服务生成日志数据。这里Facebook使用开源日志收集系统，它可以将数以百计的日志数据集存储在NFS服务器上，且大部分日志数据会复制到同一个中心的HDFS实例中，而HDFS存储的数据都会放到利用Hive构建的数据仓库中。Hive提供了类SQL的语言来与MapReduce结合，创建并发布多种摘要和报告，以及在它们的基础上进行历史分析。Hive上基于浏览器的接口允许用户执行Hive查询。Oracle和MySQL数据库用来发布摘要，这些数据容量相对较小，但查询频率较高，并需要实时响应。

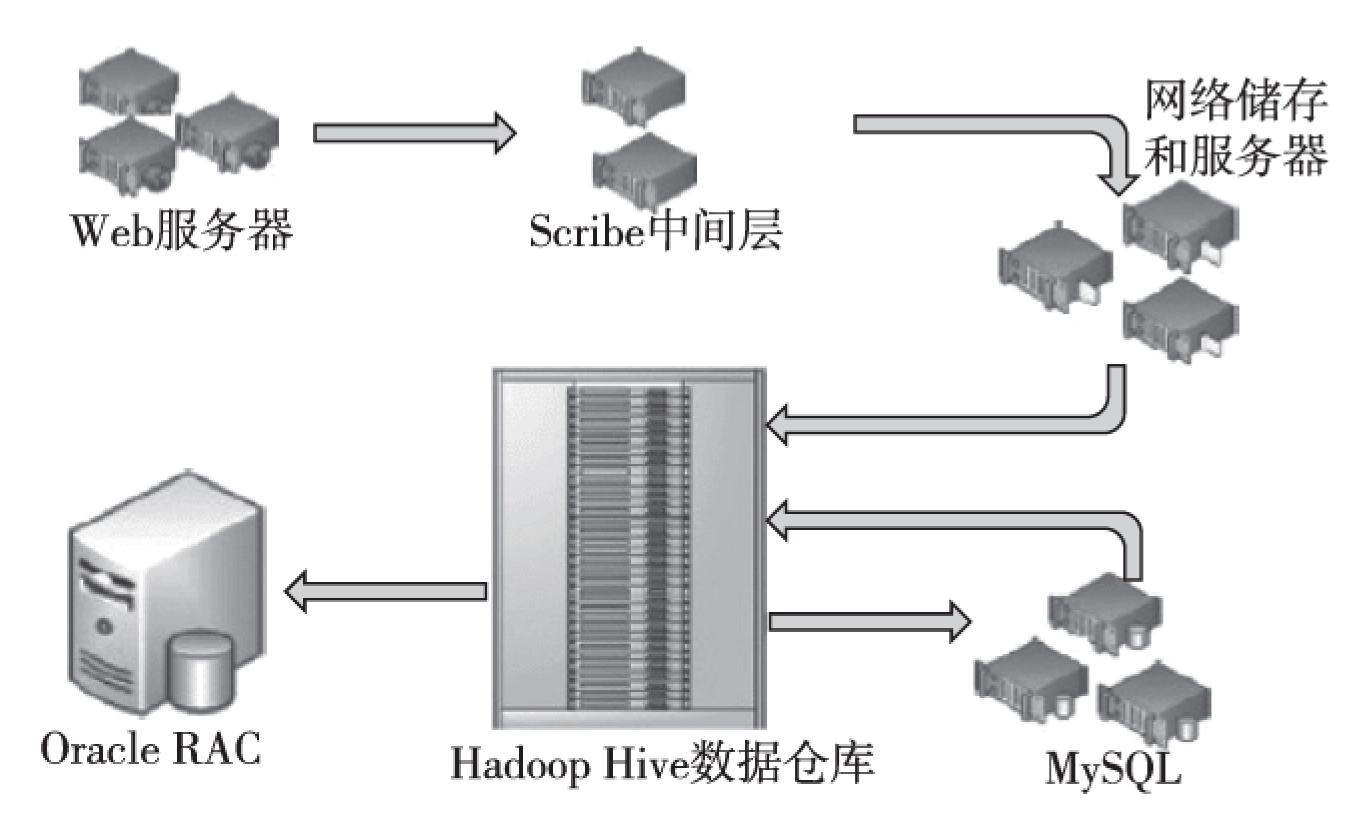


图4.50　Facebook数据仓库架构

4. AvatarNode和调度策略

一些旧的数据需要及时归档，并存储在较便宜的存储器上，如图4.51所示。下面介绍Facebook在AvatarNode和调度策略方面所做的一些工作。AvatarNode主要用于HDFS的恢复和启动。若HDFS崩溃，按原有技术恢复首先需要花10～15分钟来读取12GB的文件镜像并写回，还要用20～30分钟处理来自2000个DataNode的数据块报告，最后用40～60分钟来恢复崩溃的NameNode和部署软件。表4.2说明了BackupNode和AvatarNode的区别。AvatarNode作为普通的NameNode启动，处理所有来自DataNode的消息。AvatarDataNode与DataNode相似，支持多线程和针对多个主节点的多队列，但无法区分原始和备份。人工恢复使用AvatarShell命令行工具，AvatarShell执行恢复操作并更新ZooKeeper的zNode，恢复过程对用户来说是透明的。分布

式Avatar文件系统实现在现有文件系统的上层。

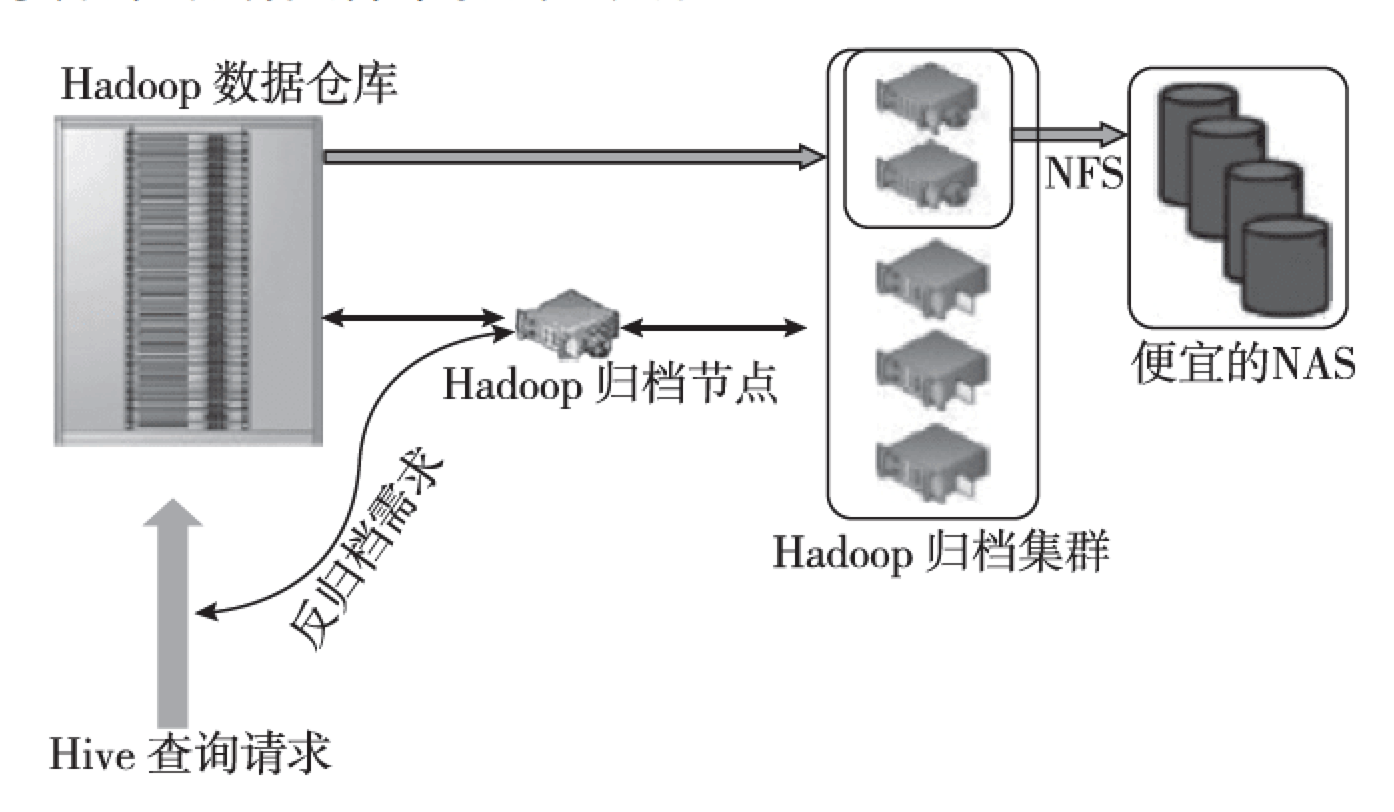


图4.51　数据归档

表4.2　BackupNode和AvatarNode的区别

BackupNode（冷备份）	AvatarNode（热备份）
Namespace状态与原始的同步 没有数据块和DataNode信息 在可用之前仍需20～30分钟	Namespace状态与原始相比有几个事务的延迟 拥有全部的数据块和DataNode信息 6500万个文件的恢复不超过一分钟

基于位置的调度策略在实际应用中存在着一些问题，如需要高内存的任务可能会被分配给拥有低内存的TaskTracker；CPU资源有时未被充分利用；为不同硬件的TaskTracker进行配置也比较困难等。Facebook采用基于资源的调度策略，即公平享有调度方法，实时监测系统并收集CPU和内存的使用情况。调度器会分析实时的内存消耗情况，然后在任务之间公平分配任务的内存使用量。它通过读取/proc/目录解析进程树，并收集进程树上所有的CPU和内存的使用信息，然后通过TaskCounters在心跳（heartbeat）时发送信息。

5. Hive的架构

Facebook的数据仓库使用Hive。这里HDFS支持三种文件格式：文本文件（TextFile），方便其他应用程序读写；顺序文件（SequenceFile），只有Hadoop能够读取并支持分块压缩；RCFile，使用顺序文件基于块的存储方式，每个块按列存储，这样有较好的压缩率和查询性能。Facebook未来会在Hive上进行改进，以使Hive支持索引、视图、子查询等新功能。

6. 挑战

现在Facebook使用Hadoop遇到的挑战如下。

- 服务质量和隔离性方面：较大的任务会影响集群性能。
- 安全性方面：如果软件漏洞导致NameNode事务日志崩溃该如何处理。
- 数据归档方面：如何选择归档数据，以及数据如何归档。
- 性能提升方面：如何有效地解决瓶颈等。

4.5　MapReduce模型概述

Hadoop的MapReduce采用的是Master/Slave架构，如图4.52所示。它主要由这样几个组件

组成：Client、JobTracker、TaskTracker 和Task。下面分别对这几个组件进行介绍。

图4.52　MapReduce结构图

1. Client

Client是用户编写的MapReduce程序与JobTracker交互的桥梁。Client提供了一些接口，利用这些接口用户可以查看作业的运行状态。MapReduce程序用“作业”来表示。每个MapReduce 程序可对应许多个作业，而每个作业会被划分成多个Map/Reduce任务。

2. JobTracker

JobTracker负责资源的监控以及作业的调度，它在主节点上单独工作。JobTracker主要是监控所有TaskTracker和作业的状态，如果发现有故障，JobTracker就会将相应的任务转分配给其他节点去完成。同时，JobTracker也会处理一些细节的工作，比如监控任务执行到哪一步了，资源使用的情况等，并将获取到的结果转达给任务调度器，让其去分配更符合的任务来调用这些资源。

3. TaskTracker

TaskTracker所做的工作是每隔一段时间就向JobTracker报告本节点上的资源使用状况、任务的调用和执行情况，同时对JobTracker 发送过来的命令进行接收并且执行相应的动作（比如开启、关闭任务等）。

4. Task

Task是由TaskTracker启动的具体任务，包括Map Task和Reduce Task。HDFS存储数据的基本单位为block，而MapReduce的处理单位是split。split里只具有一些基本的数据信息，比如数据的长度、起始位置、所属节点等。用户可以自行决定其分解方法。由于每个split由一个Map Task负责，因此split越多，Map Task也越多。

4.5.1 Map和Reduce函数

前面提到过，用户只需要根据任务要求编写MapReduce中的Map()和Reduce()这两个函数，即可完成简单的分布式程序设计。以下是MapReduce中这两个过程的公式表达。

Map过程：Map(key1,value1) → list(key2,value2)

Reduce过程：Reduce(key2, list (value2)) → list(value3)

下面用图4.53来帮助读者理解MapReduce原理。

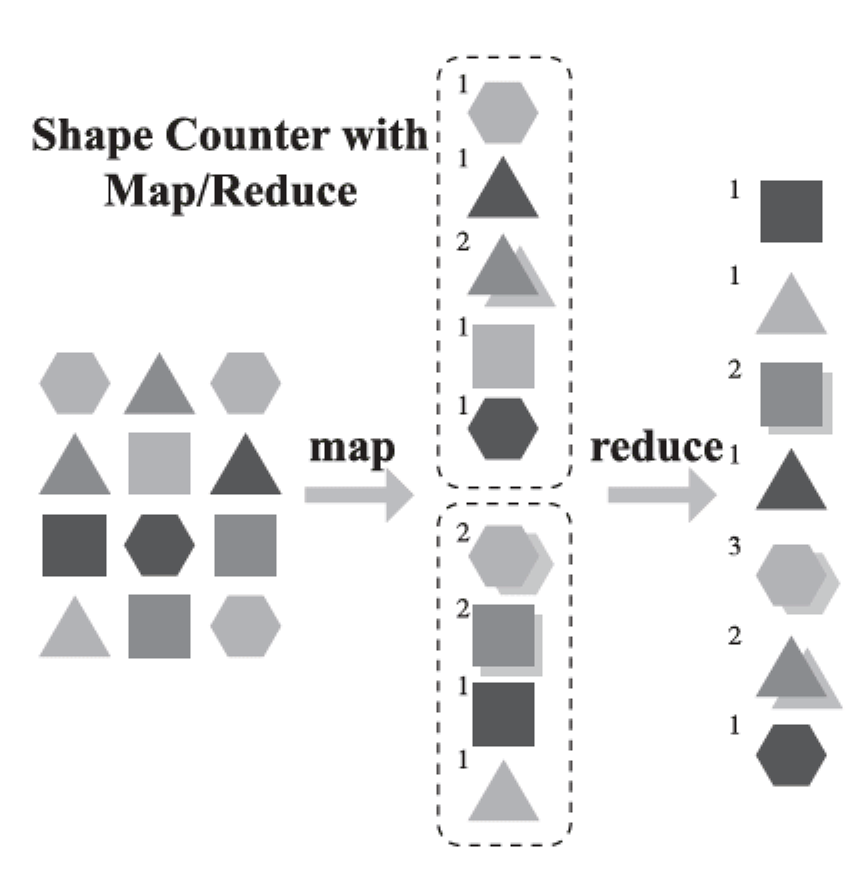


图4.53　MapReduce原理示意图

4.5.2 MapReduce工作流程

MapReduce在大数据处理中有着举足若轻的地位，那么其工作流程是怎样的呢？图4.54为MapReduce工作流程示意图。

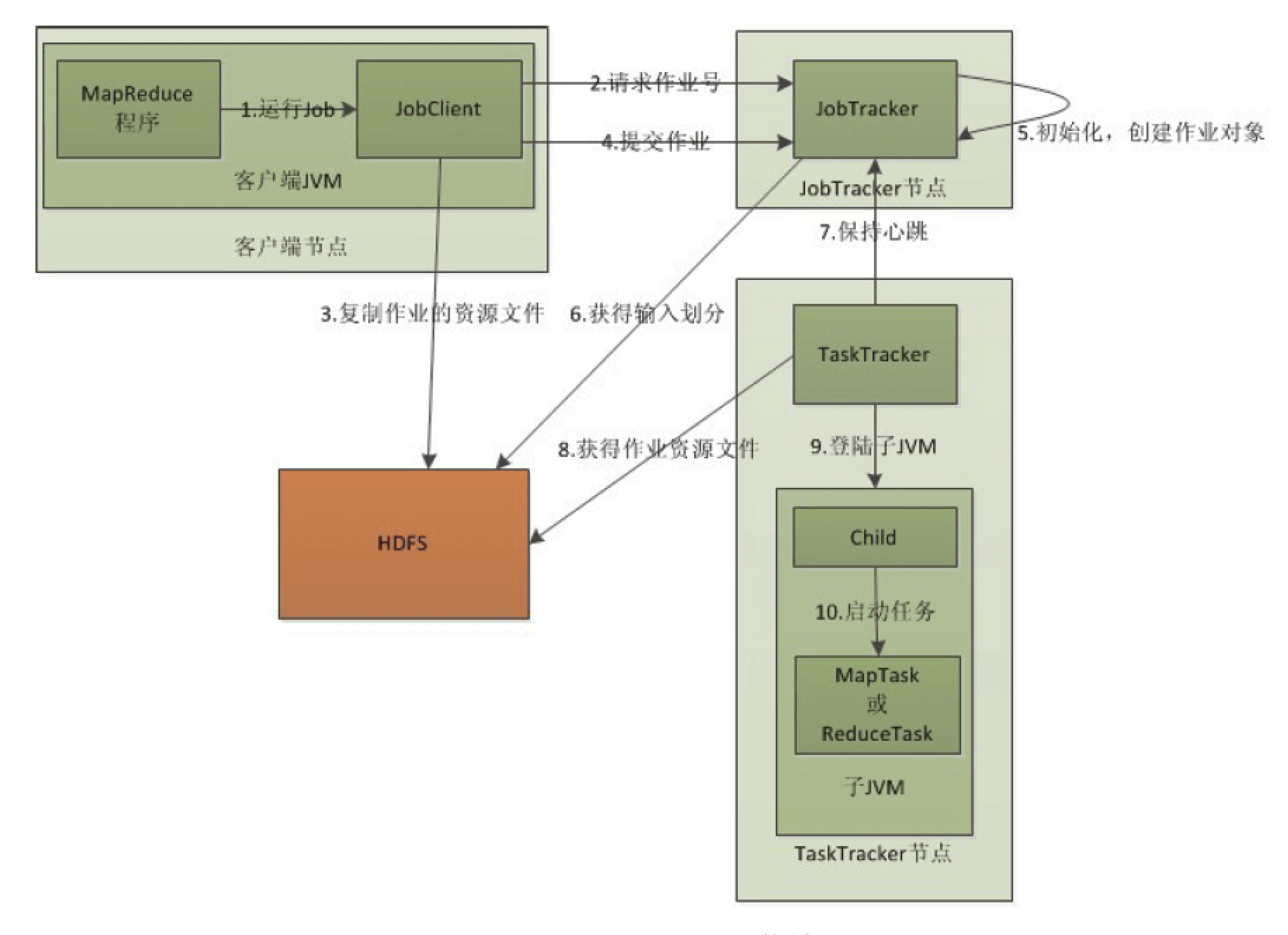


图4.54　MapReduce工作流程图

1. 工作流程概述

图4.54展示了MapReduce的工作流程，该流程的主要处理步骤如下。

（1）首先，在Client处启动一个job。

（2）向JobTracker请求为作业分配一个ID。

（3）复制作业的资源文件到HDFS上，这些资源包括MapReduce程序打包的JAR文件、配置文件和客户端计算所得的输入划分信息。这些文件都能在JobTracker专门为该作业创建的文件夹中找到。文件夹的名字是该作业的Job ID。

（4）JobTracker在接收到作业后，将该作业放在一个作业队列里，并且等待作业调度器对其进行调度。当作业调度器调度到该作业时，会依照输入划分信息，为每个划分的部分建立一个Map任务，并将Map任务交付给TaskTracker执行。

（5）TaskTracker每隔一段时间会给JobTracker发送一个Heartbeat（心跳），告诉JobTracker它的状态，同时在Heartbeat中还携带着很多诸如Map任务完成的进度等信息。

（6）在JobTracker收到job的最后一个Map/Reduce任务完成的消息后，就会把该job的状态更改为“成功”。当JobClient进行作业查询的状态时，即可发现该任务已经完成。

2. MapReduce各个执行阶段

通常说来，Hadoop的一个简单的MapReduce任务，执行的各个阶段流程和所用到的各部分功能如图4.55所示。该流程的主要处理步骤如下。

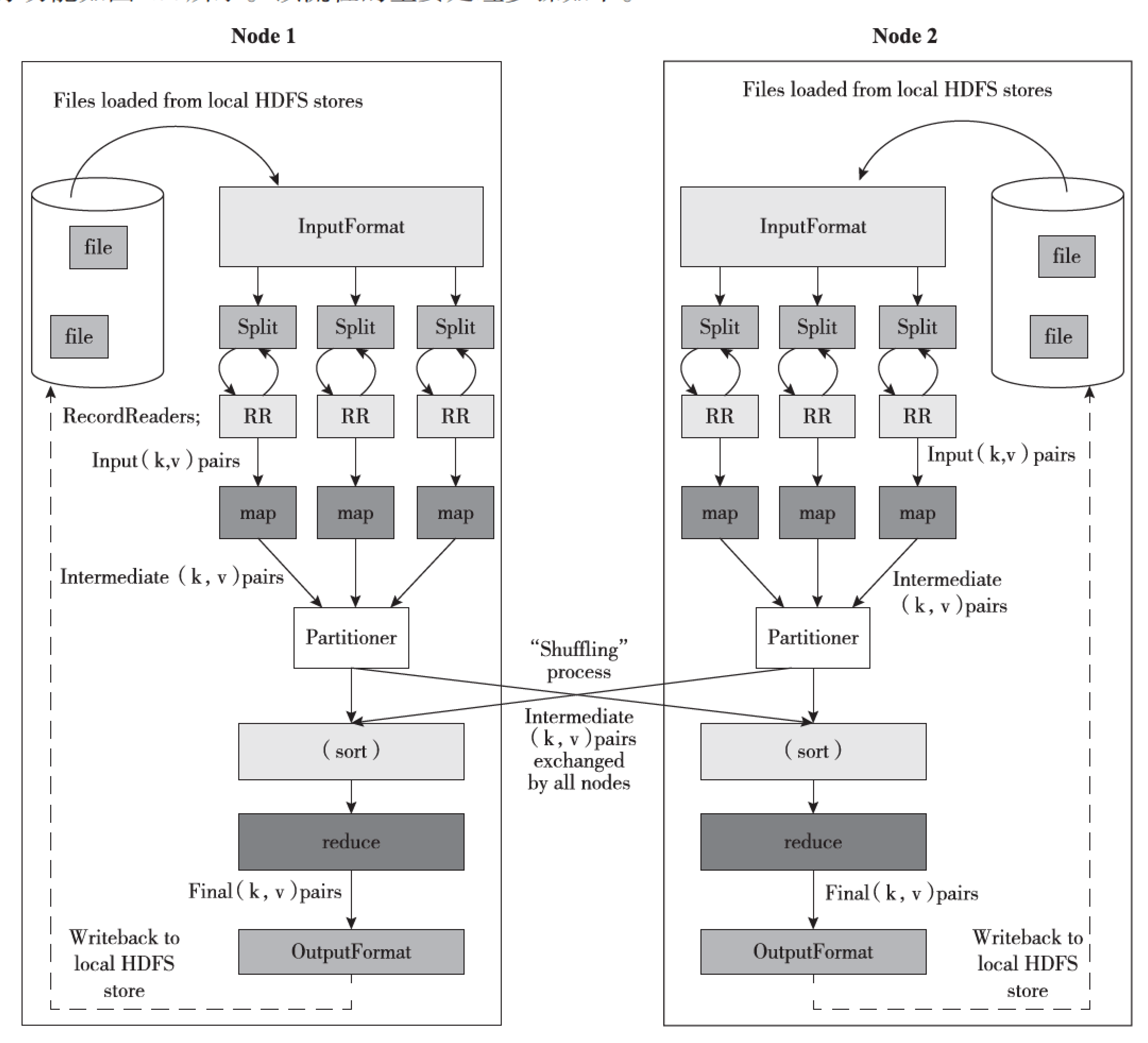


图4.55 MapReduce各个阶段流程图

（1）JobTracker在分布式环境中负责客户端对任务的建立和提交。

（2）InputFormat模块主要为Map做预处理。

（3）RecordReader处理后的结果作为Map的输入，然后Map执行定义的Map逻辑，输出处

理后的key/value对到临时中间文件。

（4）Shuffle&Partitioner，这两部分的功能主要是负责对输出的结果进行排序、分割和配置。在MapReduce流程中，为了让Reduce可以并行处理Map结果，必须由Shuffle对Map的输出进行一定的排序和分割处理，然后再交给对应的Reduce。Partitioner为Map的结果配置相应的Reduce，当Reduce很多的时候比较实用，因为它会分配Map的结果给某个Reduce进行处理，然后输出其单独的文件。

（5）Reduce处理实际的任务，得到结果，并且将结果传递给OutputFormat。

（6）OutputFormat用于测试是否已有输出目录，以及测试输出结果的类型是否属于Config中配置类型，若成立则输出Reduce汇总后的结果。

3. Shuffle过程详解

Shuffle过程是MapReduce工作流程的关键，它属于不断被优化和改进的代码库的一部分。认识Shuffle对于学习MapReduce是很重要的。

在Map和Reduce的两端之间都有Shuffle过程，它负责对数据从Map task输出到Reduce task输入的这段过程进行解释。

（1）Map端的Shuffle过程

在Map端的Shuffle过程主要是对Map的结果进行处理，分为三个步骤，分别是：partition（划分）、sort（排序）和spill（溢写），然后合并属于同一个划分的输出，再写到磁盘上，同时根据不同的划分将结果发送给相应的Reduce（Map输出的划分与Reduce的对应关系由JobTracker确定）。Reduce端又会将各个Map送来的属于同一个划分的结果合并（merge），接着对合并的结果进行排序，最后输出给Reduce进行处理，如图4.56所示。

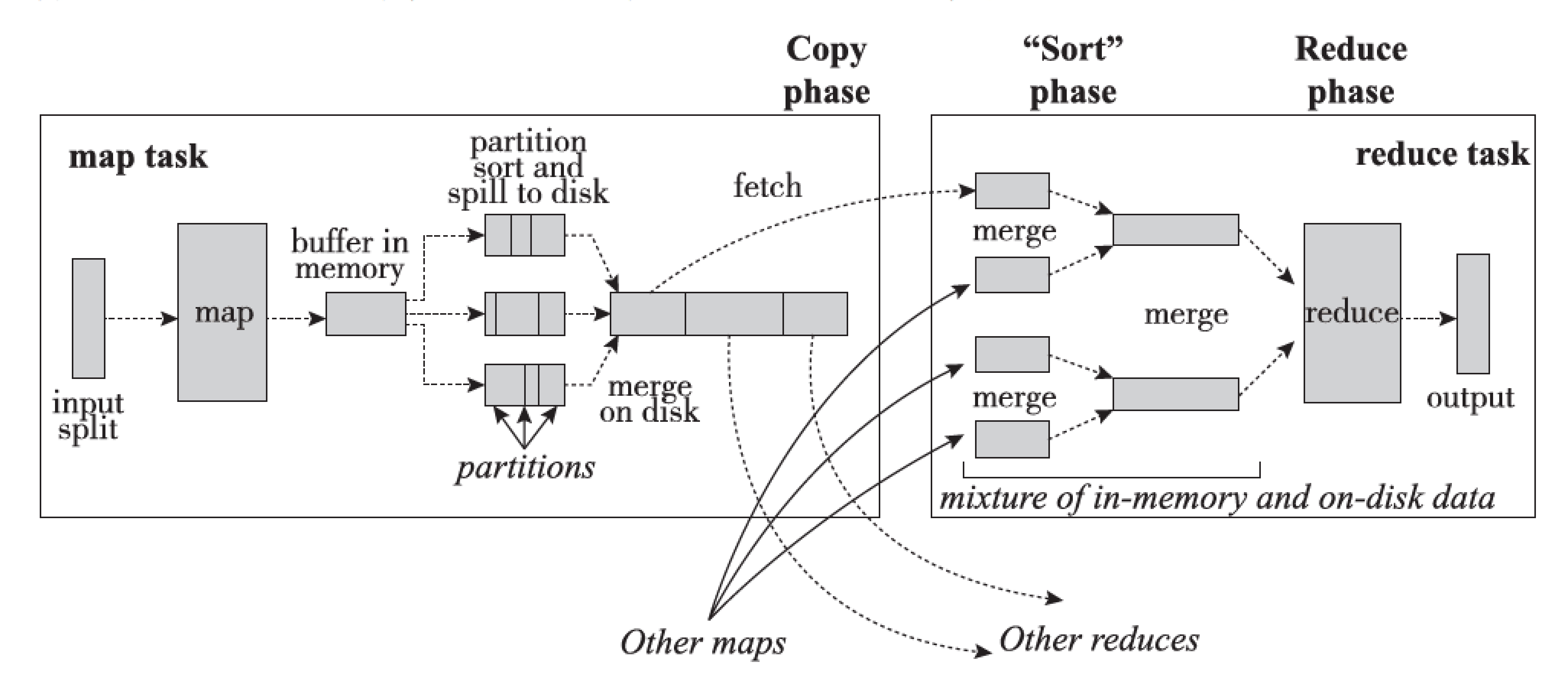


图4.56　Map端的Shuffle过程图

为了清晰地说明划分、排序与合并在MapReduce工作流程中所处的位置，读者可以通过如图4.57所示的对某个假想的Map task的运行情况来理解。

通常在每个Map task中都有一个内存缓冲区，用于存放Map的输出结果，然而当内存缓冲区不足时，需要将缓冲区的数据先放在磁盘中，以一个临时文件的方式保存，等到整个Map task结束后，再将磁盘中这个Map task产生的所有临时文件取出来进行合并，得到最后的正式输出文件，并等待Reduce task来获取数据。

Map端的Shuffle可分为四个过程：

① Map task执行

Map task从HDFS的Block中获取输入数据。

② Mapper运行

Mapper输出一个key/value对。

③ Spill

因为内存缓冲区具有一定的大小限制，所以如果Map task的输出有很多的时候，内存可能会发生溢出的情况。因此，在发生这种情况之前需要将缓冲区中的数据临时写入到磁盘，以腾出足够的空间来利用。把这个过程称之为Spill（溢写）。

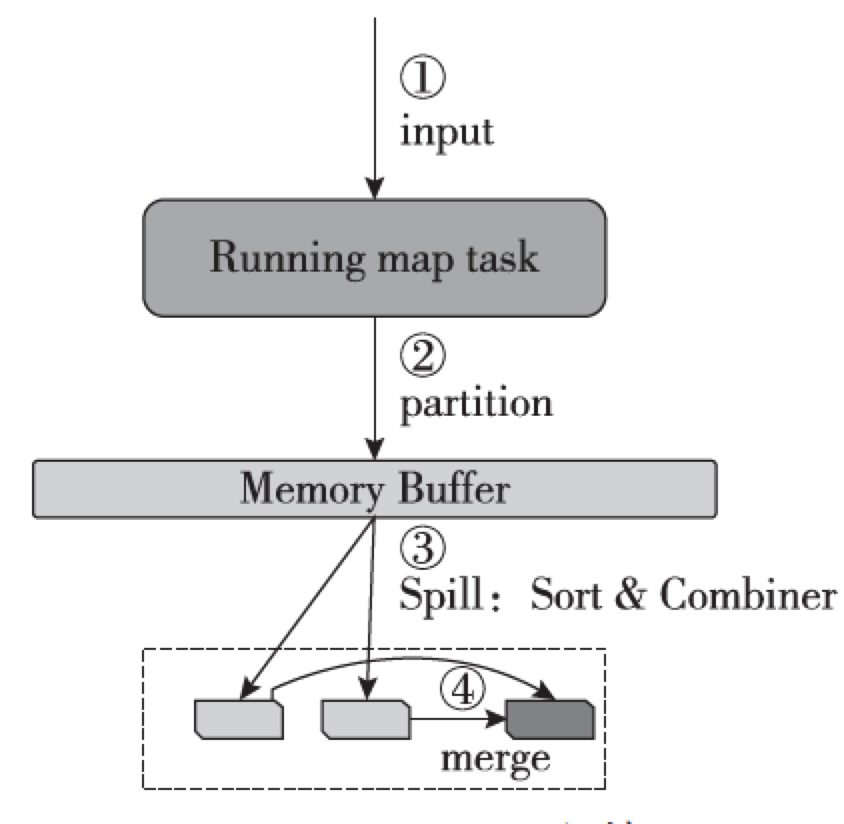


图4.57 Map task运行情况图

④ Merge

因为每次溢写都会生成一个溢写文件，所以，如果Map的输出结果很多时，磁盘上的溢写文件就会有很多。以上只是Map的中间过程，当Map task真正完成时，全部的溢写文件将会在磁盘中合并成为一个。因为最终只有一个溢写文件存在，所以将这些溢写文件归并到一起的过程就叫做Merge。

当Map task结束后，开始Reduce端的Shuffle过程。

（2）Reduce端的Shuffle过程

如果当前的Reduce 想要执行复制数据的操作，它要从JobTracker那知道已经结束的Map task。Reduce在运行之前的准备工作是获取数据和合并，这个过程不会停止。

Reduce端的Shuffle有3个过程，如图4.58所示。

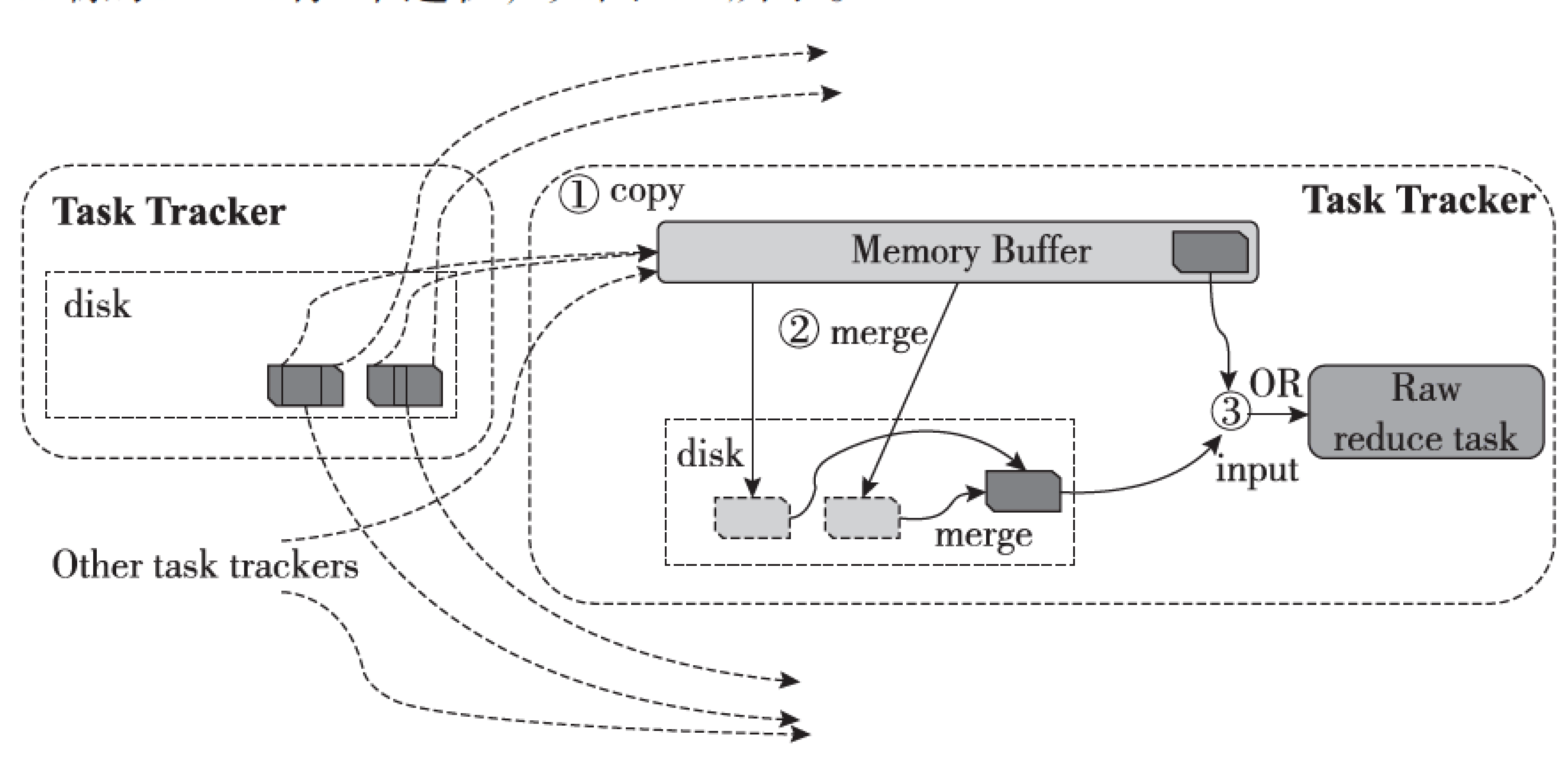


图4.58 Reduce端的Shuffle过程图

① copy阶段

主要任务是获取数据。Reduce会接收到不同Map任务传来的有序的数据。Reduce进程启动一些数据copy线程（Fetcher），通过HTTP方式请求Map task所在的TaskTracker来获取Map task的输出文件。因为Map task在Shuffle的前一个过程中早已结束，所以在本地磁盘中的这些文件就由TaskTracker来管理。

② merge阶段

Reduce端的merge与Map端的merge操作相同，不一样的是，数组中存放的是不同Map端copy来的数值。类似地，当溢写文件增多时，后台线程就会启动操作，将它们合并成为一个更大的、有序的文件，这样便能给后面的合并节省时间，提高了效率。所以读者能够看出，无论是在Map端，亦或是在Reduce端，MapReduce主要做的工作都是排序以及合并。因此常常有人说：排序是Hadoop的灵魂。

③ reducer输入文件阶段

MapReduce会将写入磁盘的数据尽量减少，这些数据是在合并的过程中生成的，而且它会将最后一次合并的结果直接传送到Reduce函数中而不是写入磁盘。在经过多次merge操作后，会在磁盘或者内存中产生一个“最终文件”。整个Shuffle最终结束的标志是确定了Reducer的输入文件是哪个。最后是执行reducer，并将结果保存到HDFS上。

简单说来，Reduce task在执行之前的工作就是不断地获取当前job里每个Map task的最终结果，然后对从不同来源的数据不断地做merge，最终形成一个文件作为Reduce task的输入文件。

4.5.3 并行计算的实现

对于一些由大量通用计算机组成的集群来说，MapReduce是能帮助它们运行并行程序的一个重要的计算模型。由集群中每一个单独的节点来执行每一个Map和Reduce任务，运算速度之快、效率之高是不言而喻的。下面介绍一下实现并行计算所需要的几个必不可少的部分。

1. 数据分布存储

回顾一下HDFS的结构（如图4.59所示）。HDFS主要是由一个NameNode和N个DataNode组成，分别扮演的是管理者和工作者的角色，每台通用计算机均用节点来表示。HDFS能够对文件进行各种各样的操作，比如创建目录，创建、复制、删除文件和查看文件的内容等。但是HDFS的底层把文件分为一个个Block，然后把这些 Block 分散地存放在不同的 DataNode 上，并且每个 Block 还可以复制成多份数据存放在不同的 DataNode 上，因此能及时恢复因灾难或者错误而丢失的数据。NameNode以维护数据结构的方式来收录每一个文件被分成了多少个Block、这些 Block 来源于哪个DataNode，以及各个 DataNode 的状态和详情等重要信息。

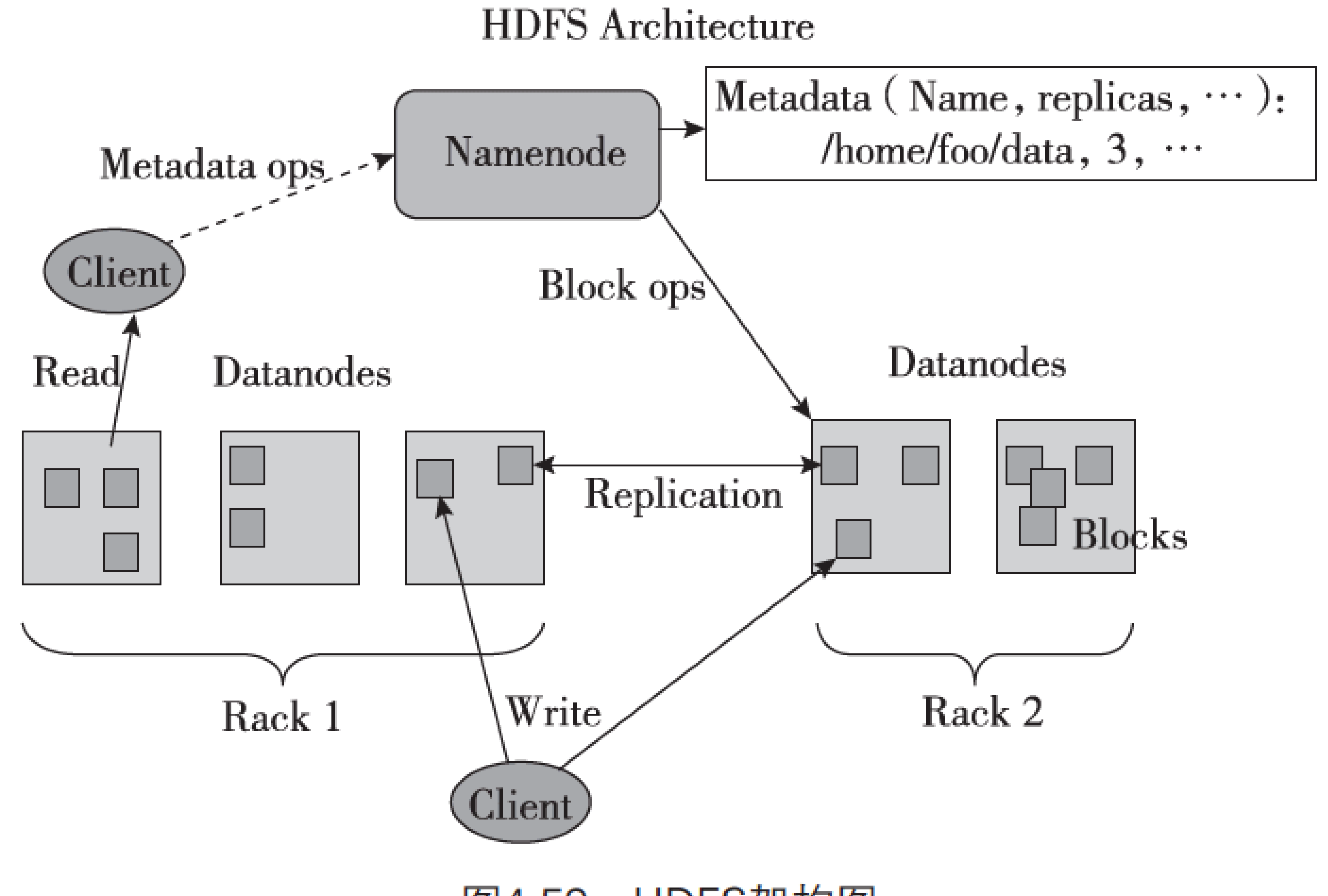


图4.59　HDFS架构图

2. 分布式并行计算

读者已经知道，JobTracker是Hadoop中一个主要的组成部分，负责调度和管理其他的TaskTracker。作为一个“管理者”，JobTracker 可以在集群中的任意一台计算机上工作。TaskTracker则担任“工作者”的角色，执行具体的任务，它必须运行于DataNode上，换言之，DataNode既有存储数据的功能，又有对数据进行计算的功能。然后，JobTracker负责把Map任务和Reduce任务分发给处于空闲状态的TaskTracker，并且使得这些任务并行执行，以及监控任务的运行情况。当某个TaskTracker出故障时，JobTracker会将其任务转交给另一个处于空闲状态的TaskTracker重新运行。

图4.60是Hadoop JobTracker的示意图。

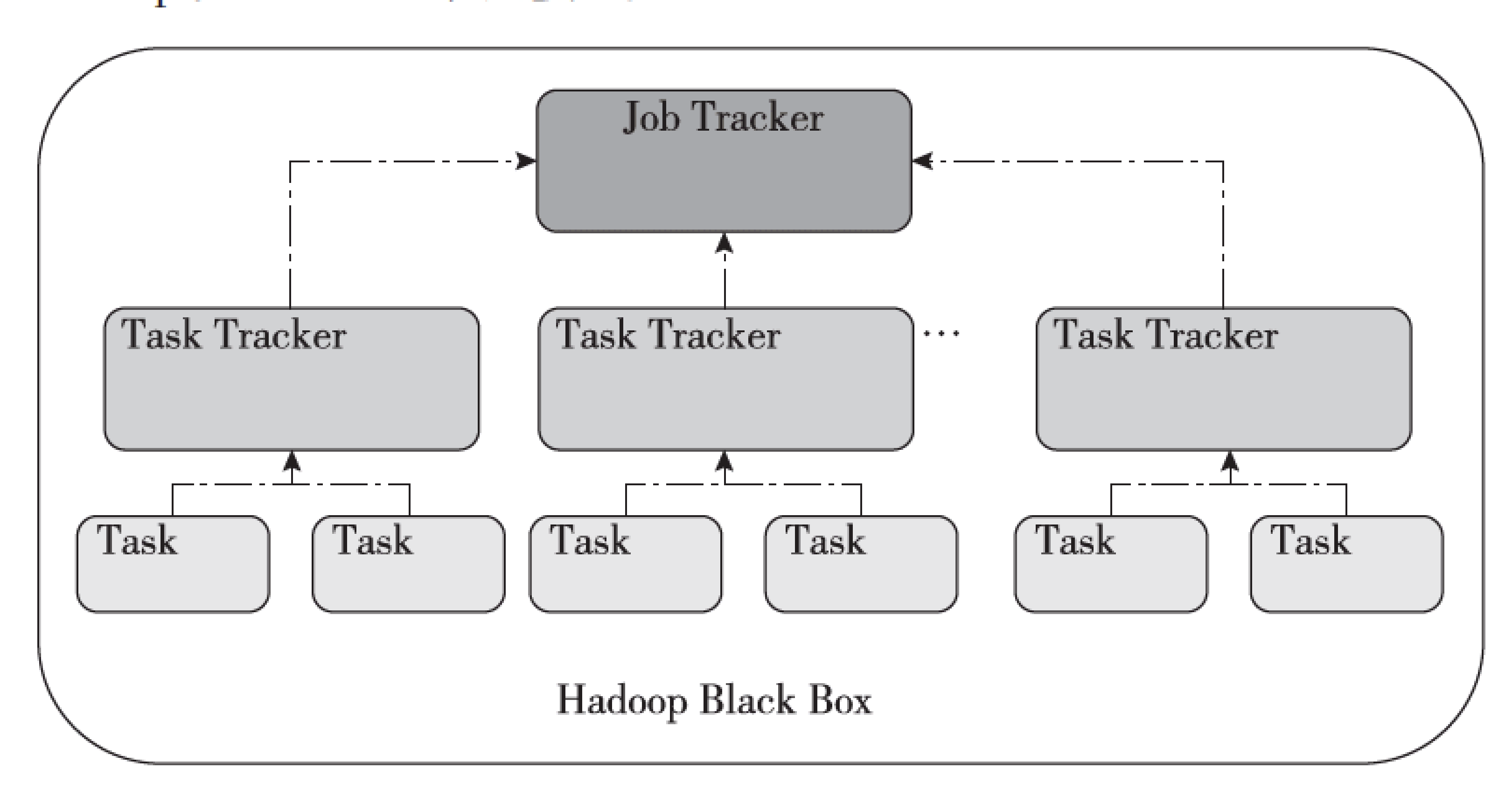


图4.60　Hadoop Job Tracker的示意图

3. 本地计算

本地计算的特点是存储数据的计算机负责计算所存储的数据，这样一来就可以大大降低网络上数据的传输次数，有效地节省带宽。Hadoop是采用集群方式的分布式并行系统，节点扩充起来比较方便，因此它可以提供强大的计算能力。但是由于数据需要在不同的计算机之间流动，故对网络带宽要求高的问题逐渐凸显出来。因为“本地计算”能够在很大程度上节约带宽资源，所以业界有说法称“移动计算比移动数据更经济”。

4. 任务粒度

在划分原始的大数据集时，通常的原则是令分割后的小数据集等于或是小于文件系统中一个数据块的大小（一般默认为64MB），如此能够确保一个独立的小数据集被分配到一台计算机上，方便本地计算任务的进行[①]。例如，如果有M个小数据集需要处理，就启动M个Map任务，注意这M个Map任务分布于N台计算机上，它们会并行运行，而用户也可以指定Reduce任务的数量R。

5. Partition

Partition是选择配置，主要功能是在多个Reduce的情况下，分配Map的结果给某个Reduce进行处理，然后输出其单独的文件。例如，Reduce任务有R个，则把Map任务输出的中间结果

① http://www.doc88.com/p-787395462005.html

按key的范围划分成R份，通常根据hash(key) mod R函数来划分。如此划分的目的主要是为了确保由同一个Reduce任务来对某一段范围之间的key进行处理，进而使Reduce过程更简单[①]。

6. Combine

在对key值进行划分前，还可以先对中间过程的结果进行合并（Combine），也就是先将中间结果中同样key值的键值对合并成一对。Combine不是必须的，但由于它的主要作用是在每一个Map执行完分析以后，在本地优先作Reduce的工作，因此可以减少在Reduce过程中的数据传输量。

由于Combine过程与Reduce过程较相似，在大部分情形下能够直接使用Reduce函数。然而Combine属于Map任务的其中一个部分，它是紧跟着Map函数执行的。Combine的加入可以减少键值对在中间结果中的数目，进而达到降低网络流量的目的。

7. Reduce任务从Map任务节点获取中间结果

在完成Partition和Combine过程后，Map任务输出的中间结果会形成文件并存放在本地磁盘中，主控JobTracker会被告知其位置。之后，JobTracker再通知Taskjob任务去DataNode上存储中间结果的位置去读数据。值得注意的是，Map任务输出所有的中间结果都将使用同一个hash函数把key值分为R份，每段key值区间由一个单独的Reduce任务负责处理。每个Reduce任务首先需要在多个Map任务节点中获取属于其负责的key值区间内的中间结果，进而执行Reduce函数，最后输出一个最终的结果文件。

8. 任务管道

如果有R个Reduce任务，就会产生R个最终结果。大部分情形下这R个结果并不需要合并成一个最终结果，因为这R个结果又可以成为另一个计算任务的输入数据，开启另一个并行计算任务，因此这个过程也就形成了任务管道。

4.6 实例分析：WordCount

如果想统计过去计算机论文中出现次数最多的几个单词是什么，可以实现的方法大致有如下几种。

（1）写一个小程序，把所有论文按顺序遍历一遍，对每一个遇到的单词出现的次数进行统计，最后就可以知道哪几个单词最热门了。这种方法在数据集比较小的情况下非常合适，且很有效，实现起来也最简单。

（2）写一个多线程程序，并发地遍历论文。理论上是可以高度并发的，因为统计一个文件时不会影响另一个文件的统计。该方法在使用多核机器或多处理器时工作效率比方法（1）高。但是，写一个多线程程序要复杂得多，开发者需要自己去同步共享数据，比如要防止两个线程重复统计文件这种情况发生。

① http://www.doc88.com/p-787395462005.html

（3） 由多台计算机完成一个作业。可以使用方法（1）的程序，在N台机器上部署，接着把论文集分成N份，每个作业由一台机器运行。这个方法运行速度足够快，但是部署起来很麻烦：既要人工把论文集分开，复制到各台机器，还要整合这N个运行结果。

（4） 使用MapReduce。MapReduce实际上就是方法（3），但对于如何拆分文件集，如何复制程序，如何整合结果这些问题的解决方案都是MapReduce框架定义好的。开发人员只要定义好这个任务（用户程序），其他工作都由MapReduce去处理。

每个拿到原始数据的计算机只需要负责切分输入的数据成为单词即可，所以单词的切分任务可以在Map阶段去完成。另外，对于相同单词的频数计算也可以进行并行化处理，方法是将相同的单词交给一台机器去统计数量，然后输出最终结果，这个任务可以交由Reduce阶段去完成。关于如何将中间结果根据不同单词分组发送给Reduce机器，这部分工作由MapReduce过程中的Shuffle去完成。图4.61所示为MapReduce进行过程示意图。

MapReduce Flow

Input = { Key，Value | Key，Value | … }
Split Input into Key–Value pairs.

Map | Map | Map
For each K–V pair call Map.

Key，Value / Key，Value / … | Key，Value / Key，Value / … | Key，Value / Key，Value / …
Each Map produces new set of K–V pairs.

Sort
Reduce（K，V[]）
For each distinct key，call reduce. Produces one K–V pair for each distinct key.

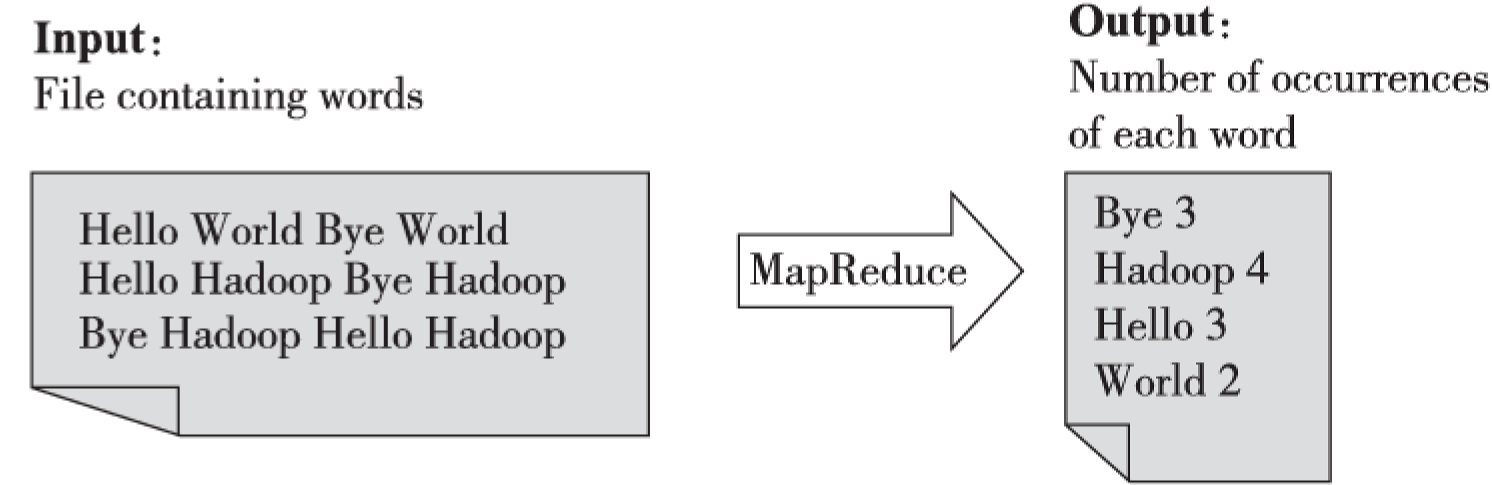
output = { Key，Value | Key，Value | … }
Output as a set of Key Value Pairs.

（a）

MapReduce WordCount Example

Input：
File containing words

Output：
Number of occurrences of each word



How can we do this within the MapReduce framework?
Basic idea：parallelize on lines in input file!

（b）

图4.61　MapReduce运行过程示意图

4.6.1 WordCount设计思路

WordCount的整个设计思路如下。

（1）Map阶段主要任务是划分输入数据为单词。

（2）Shuffle阶段对相同的单词进行聚集和分发（这个过程属于MapReduce的默认过程，不需要具体配置）。

（3）Reduce阶段接收所有的单词并且统计各单词的频数。

由于MapReduce中传递的数据都是<key, value>形式的，并且Shuffle排序聚集分发是按照key值进行的，因此，将Map的输出设计成由word表示key，1表示value的形式。它表明单词出现了1次（Map的输入方式为Hadoop默认方式，即文件的一行表示value，行号表示key）。

Reduce输入的是Map输出聚集后的结果，即<key, value-list>，对于本实例就是<word, {1,1,1,1,…}>。Reduce输出的是与Map输出相同的形式，只是后面的数值不再是固定的1，而是具体算出的word所对应的数量。

经历过程如下。

Input

```
Hello World Bye World
Hello Hadoop Bye Hadoop
Bye Hadoop Hello Hadoop
```

Map

```
<Hello,1>
<World,1>
<Bye,1>
<World,1>
<Hello,1>
<Hadoop,1>
<Bye,1>
<Hadoop,1>
<Bye,1>
<Hadoop,1>
<Hello,1>
<Hadoop,1>
```

Sort

```
<Bye,1>
<Bye,1>
<Bye,1>
<Hadoop,1>
<Hadoop,1>
<Hadoop,1>
```

```
<Hadoop,1>
<Hello,1>
<Hello,1>
<Hello,1>
<World,1>
<World,1>
```

Combine

```
<Bye,1,1,1>
<Hadoop,1,1,1,1>
<Hello,1,1,1>
<World,1,1>
```

Reduce

```
<Bye,3>
<Hadoop,4>
<Hello,3>
<World,2>
```

4.6.2　WordCount代码[1]

WordCount代码具体如下所示。

```
package com.felix;
import java.io.IOException;
import java.util.Iterator;
import java.util.StringTokenizer;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;
```

① http://www.docin.com/p-144322396.html.

```
import org.apache.hadoop.mapred.TextInputFormat;
import org.apache.hadoop.mapred.TextOutputFormat;
public class WordCount
{
 /**
* MapReduceBase类：实现了Mapper和Reducer接口的基类（其中的方法只是实现接口，而未作任何事情）
* Mapper接口：
* WritableComparable接口：实现WritableComparable的类可以相互比较。所有被用作key的类应该实现此接口。
* Reporter可用于报告整个应用的运行进度，本例中未使用。
*
*/
public static class Map extends MapReduceBase implements
        Mapper<LongWritable, Text, Text, IntWritable>
    {
        /**
       * LongWritable, IntWritable, Text均是Hadoop中实现的用于封装 Java 数据类型的类，这些类实现了WritableComparable接口。
       * 都能够被串行化从而便于在分布式环境中进行数据交换，读者可以将它们分别视为long，int和String 的替代品。
         */
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();
        /**
        * Mapper接口中的Map方法：
        * void map(K1 key, V1 value, OutputCollector<K2,V2> output, Reporter reporter)
         * 映射一个单个的输入k/v对到一个中间的k/v对
         * 输出对不需要和输入对类型相同，输入对可以映射到0个或多个输出对。
         * OutputCollector接口：收集Mapper和Reducer输出的<k,v>对。
         * OutputCollector接口的collect(k, v)方法：增加一个(k,v)对到output
         */
        public void map(LongWritable key, Text value,
               OutputCollector<Text, IntWritable> output, Reporter reporter)
               throws IOException
        {
```

```
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);
            while (tokenizer.hasMoreTokens())
            {
                word.set(tokenizer.nextToken());
                output.collect(word, one);
            }
        }
    }
    public static class Reduce extends MapReduceBase implements
            Reducer<Text, IntWritable, Text, IntWritable>
    {
        public void reduce(Text key, Iterator<IntWritable> values,
               OutputCollector<Text, IntWritable> output, Reporter reporter)
               throws IOException
        {
            int sum = 0;
            while (values.hasNext())
            {
                sum += values.next().get();
            }
            output.collect(key, new IntWritable(sum));
        }
    }
    public static void main(String[] args) throws Exception
    {
        /**
 * JobConf: map/reduce的job配置类，向Hadoop框架描述map-reduce执行的工作
              * 构造方法：JobConf()、JobConf(Class  exampleClass)、JobConf
                (Configuration conf)等。
              */
        JobConf conf = new JobConf(WordCount.class);
        conf.setJobName("wordcount");            //设置一个用户定义的job名称
        conf.setOutputKeyClass(Text.class);     //为job的输出数据设置Key类
        conf.setOutputValueClass(IntWritable.class);   //为job输出设置value类
        conf.setMapperClass(Map.class);          //为job设置Mapper类
        conf.setCombinerClass(Reduce.class);       //为job设置Combiner类
```

```
        conf.setReducerClass(Reduce.class);         //为job设置Reduce类
        conf.setInputFormat(TextInputFormat.class);     //为map-reduce任务设置
                                                        InputFormat实现类
        conf.setOutputFormat(TextOutputFormat.class);   //为map-reduce任务设置
                                                        OutputFormat实现类
        /**
         * InputFormat描述map-reduce中对job的输入定义
         * setInputPaths(): 为map-reduce job设置路径数组作为输入列表
         * setInputPath(): 为map-reduce job设置路径数组作为输出列表
         */
        FileInputFormat.setInputPaths(conf, new Path(args[0]));
        FileOutputFormat.setOutputPath(conf, new Path(args[1]));
        JobClient.runJob(conf);         //运行一个job
    }
}
```

4.6.3　过程解释

Map操作的是将输入转换成<key,value>的形式，其中，key是文档中某行的行号，value是该行的内容。Map操作会将输入文档中每一个单词的出现输出到中间文件中去，如图4.62所示。

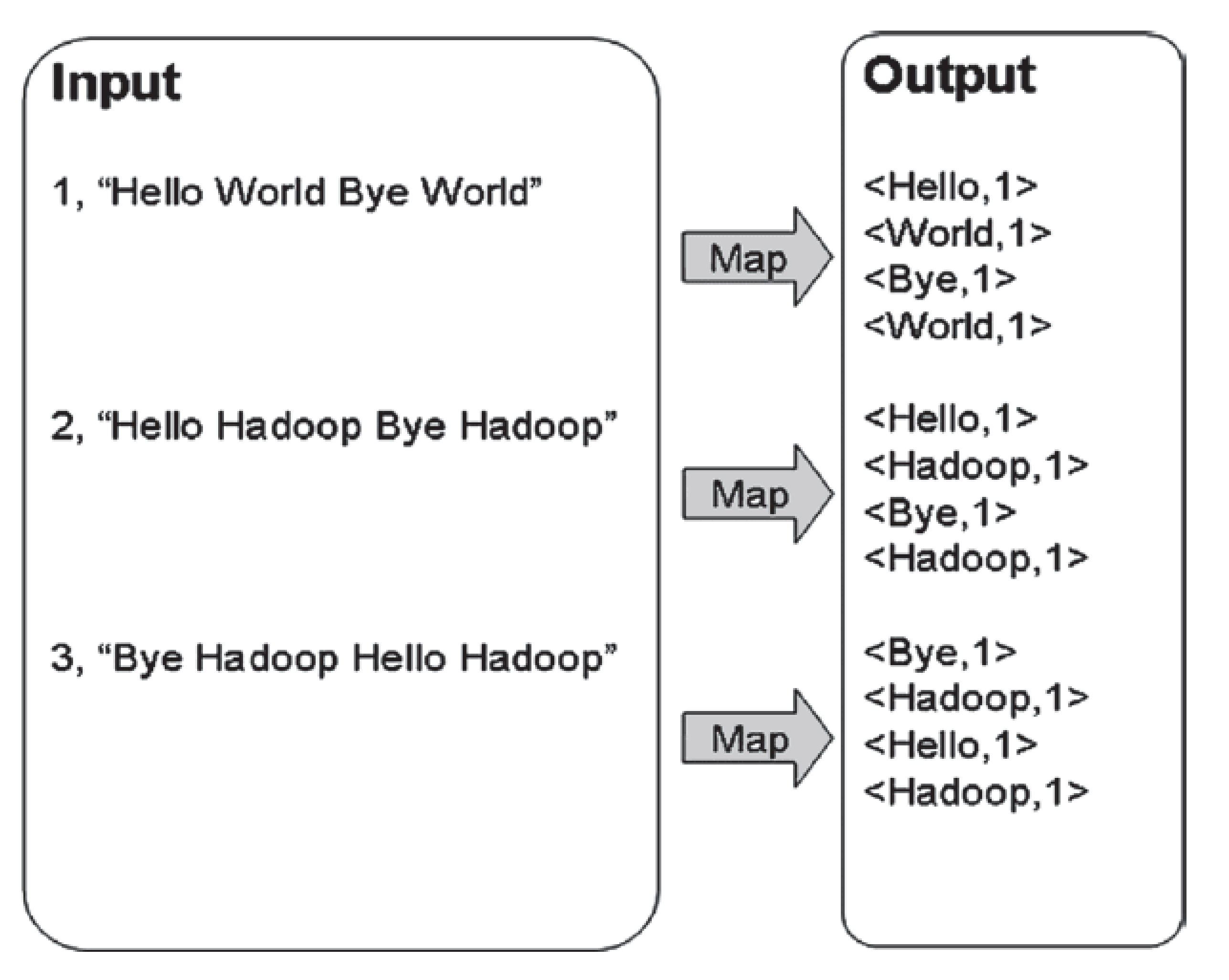


图4.62　Map操作过程示意图

Reduce操作的是输入的单词和该单词出现次数的序列，如<"Hello", [1,1,1]>，< "World", [1,1] >，< "Bye", [1,1,1] >，< "Hadoop", [1,1,1,1] >等。然后根据每个单词，算出总的出现次数。最终的结果是：< "Bye",3>，< "Hadoop",4> ，< "Hello",3>，< "World",2>，如图4.63所示。

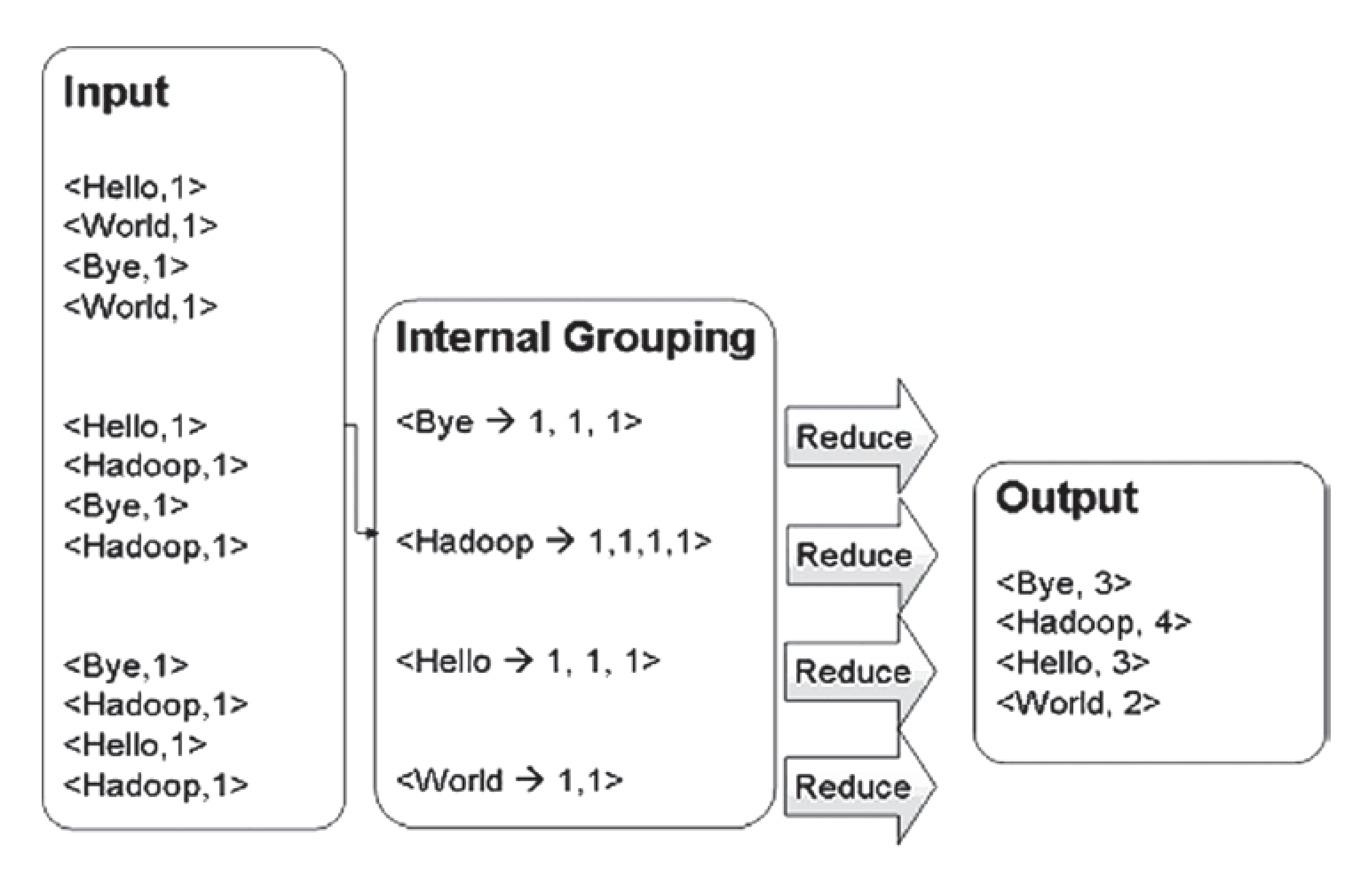


图4.63 Reduce操作过程示意图

整个MapReduce过程具体的运行流程如下。

（1）在MapReduce Library中将Input分成M份。

（2）Master把M份job分配给处于空闲状态的M个worker来处理。

（3）对于每一个正处于被输入状态的<key，value>进行Map操作，然后将中间结果作为缓冲保存在内存中。

（4）每隔一段时间（或者根据内存状态）将缓冲区中的中间信息重新写到本地磁盘上，同时把文件信息传回给Master（Master需要把这些文件信息传递给Reduce worker）。

（5）R个Reduce worker同时开始工作，从不同的Map worker的Partition那里获得数据，接着用key进行排序。

（6）由Reduce worker对中间数据进行遍历，并且对于每个惟一的Key，执行Reduce函数（其参数是该key以及其相关的一系列value）。

（7）执行完毕后，对用户程序进行唤醒，并且返回结果（最后的Output应该有R份，每个Reduce Worker一个）。

就上面的实例来说，由于每个文档中都可能会产生许多的诸如（“the”，1）这样的中间结果，因此琐碎的中间文件很大程度上会导致传输上的损失。所以，MapReduce提供了Combiner函数来支持用户。这个函数通常与Reduce函数实现的功能相同，不同的地方是Reduce函数的输出是最终结果，而Combiner函数输出的则是作为Reduce函数中某一个输入的中间文件，如图4.64所示。

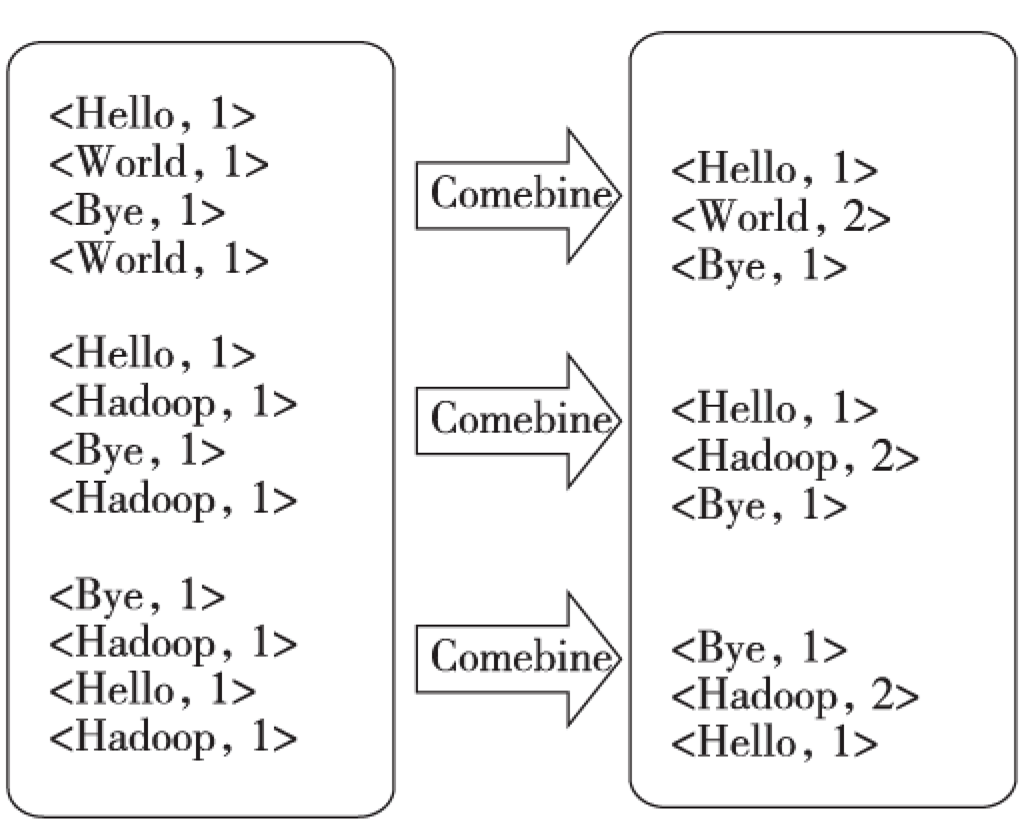


图4.64 Combiner函数过程示意图

图4.65所示是以Combiner输出结果作为输入的Reduce过程示意图。

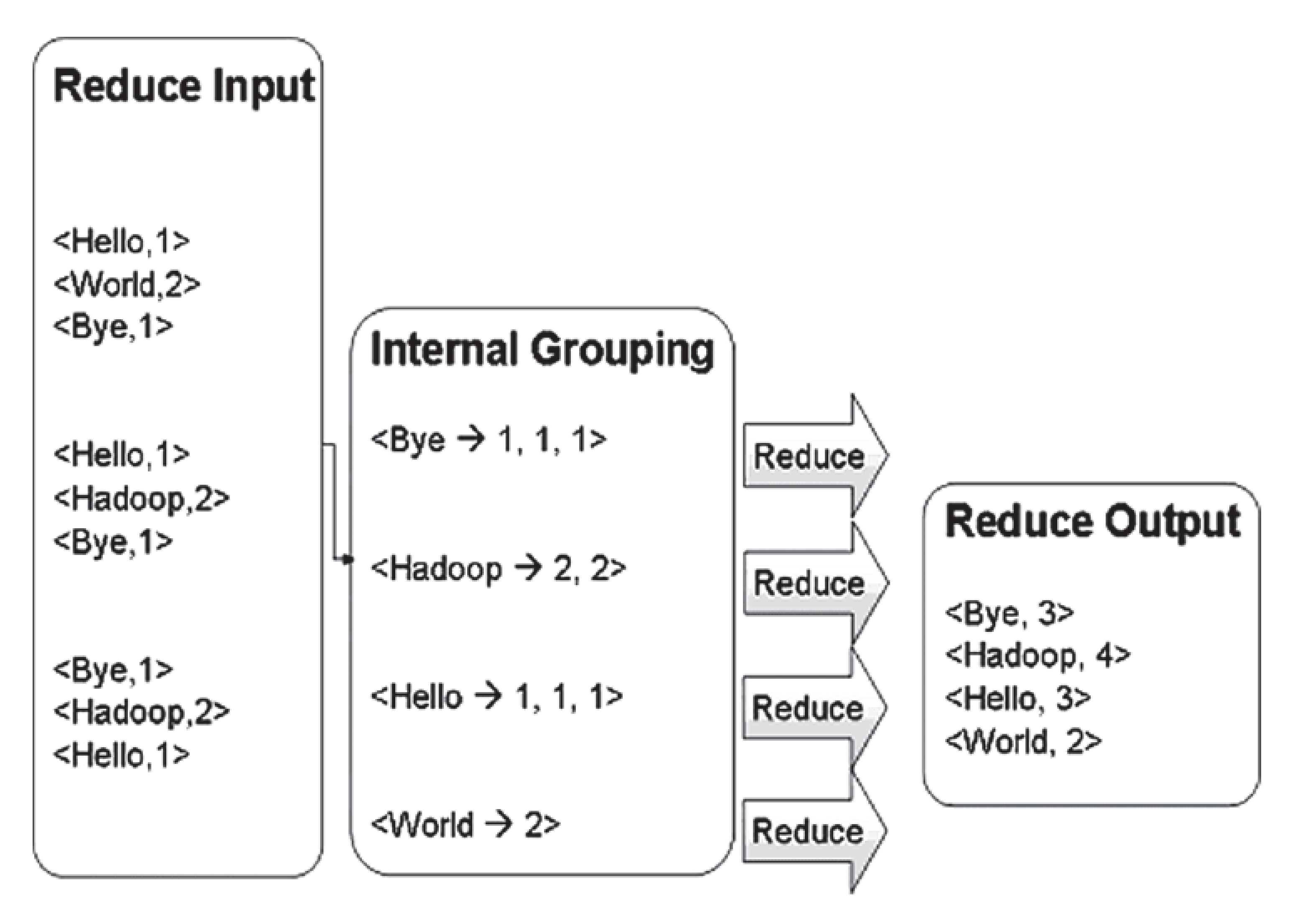


图4.65　以Combiner输出结果作为输入的Reduce过程示意图

4.7　练习

1. Hadoop的定义是什么？简要说明Hadoop的特点。
2. 在Hadoop的体系结构中两个核心的组件是什么？
3. 简述MapReduce的定义以及其中包含的两个函数Map()和Reduce()的原理。
4. MapReduce的工作流程有哪几步？
5. 你还能想到其他关于Hadoop的应用实例吗？

参考文献

[1] 蔡斌，陈湘萍. Hadoop技术内幕：深入解析Hadoop Common和HDFS架构设计与实现原理[M]. 北京：机械工业出版社，2013.

[2] Lam C. Hadoop实战[M]. 北京：人民邮电出版社，2011.

[3] 李豪. 采用MapReduce与Hadoop进行大数据分析[M/OL]. 2014.

[4] 陆嘉恒. HADOOP实战[M]. 北京：机械工业出版社，2011.

[5] White T. Hadoop权威指南第2版[M]. O'Reilly Media，Inc，2011.

[6] Hadoop技术论坛. 在Windows上安装Hadoop教程[M/OL]，2014.

[7] 陆嘉恒. Hadoop应用案例分析：在Facebook的应用[M/OL]，2014.

[8] 林子雨. 大数据技术基础[M/OL]，2013.

[9] 一见. 在Windows上安装Hadoop教程[J/OL]. Hadoop开发者2010年入门专刊，2010.

[10] 小米. 在Linux上安装Hadoop教程[J/OL]. Hadoop开发者2010年入门专刊，2010.

[11] 王承才. 小学校园Web网络硬盘应用系统的研究及实现[D]. 广州：华南理工大学，2011.

[12] 孙洪波. OSS/BSS云部署中的分布式计算特点的研究及应用[D]. 南京：南京邮电大学，2013.

[13] 朱贤军. 基于物联网技术的餐饮业油烟实时监测移动平台[D]. 淮南：安徽理工大学，2013.

[14] 喻承，杨树强，肖英. 面向海量数据非关系数据库的测试基准研究[A]. 第九届中国通信学会学术年会论文集[C]. 北京：北京邮电大学出版社，2012.

[15] 陈浩. 基于Hadoop的微博用户影响力排名算法研究[D]. 上海：华东理工大学，2013.

[16] 林纪坡. 基于Hadoop的高性能文本聚类算法的设计与实现[D]. 兰州：西北师范大学，2013.

[17] 李步源. 基于云计算的协同过滤算法并行化研究[D]. 郑州：郑州大学，2013.

[18] 胡志刚. 基于协同的并行设计环境理论与方法研究[D]. 长沙：中南大学，2002.

[19] me115. Hadoop权威指南–中文版（前三章）[EB/OL]. 2010 .

[20] 抚苏. 大数据带来价值[N]. 电脑报. 2013–07–08（014）.

[21] 李雅琼. 基于weka的web文本挖掘的研究和实现[D]. 郑州：郑州大学，2013 .

[22] 谭洁清，毛锡军. Hadoop云计算基础架构的搭建和HBase和hive的整合应用[J]. 贵州科学. 2013，31（5）：32–35.

[23] 小鱼儿. 使用Cygwin模拟Linux环境安装配置运行基于单机的Hadoop[EB/OL]，2011.

[24] quqi99. Hadoop知识分享文稿[EB/OL]，2011.

[25] 西电一枝花. hadoop hdfs搭建mapreduce环境搭建wordcount程序简单注释[EB/OL]. 2011.

[26] muyannian. Hadoop MapReduce教程[EB/OL]. 2010.

[27] 猶大之死. Linux平台下Hadoop的安装配置[EB/OL]. 2014

[28] Sweblish. 基于mapreduce的Hadoop join实现分析（一）[EB/OL]. 2012.

[29] 吹之旅. 三网融合及承载业务[EB/OL]. 2012.

[30] zeh_lm. 基于mapreduce的Hadoop join实现分析[EB/OL]. 2010.

[31] 晓敏. Hadoop[EB/OL]. 2012.

[32] 余楚礼. 基于Hadoop的并行关联规则算法研究[D]. 天津：天津理工大学，2011.

[33] 许伟静. 云计算在媒体资源管理系统中的应用研究[D]. 北京：北京化工大学，2013.

[34] whlinuxlover. 海量数据处理与存储调研[EB/OL]. 2012.

[35] 雨落. hadoop之wordcount分析[EB/OL]. 2011.

第5章

数据查询和分析的高级技术

数据查询和数据分析是实现数据应用的两个关键步骤，本章分别介绍了数据查询和数据分析的方法和技术。数据查询部分，详细介绍了SQL on Hadoop查询技术，其中包括Hive查询技术、实时交互SQL查询以及基于PostgreSQL的SQL on Hadoop查询技术等内容。数据分析部分，详细介绍了数据基本分析方法、高级分析方法以及可视化技术，并介绍了一些数据分析工具，比如统计分析工具、数据挖掘工具以及可视化设计工具等。

5.1　SQL on Hadoop查询技术

目前大部分的大数据存储都不是基于关系型数据库的，所以用传统的SQL语言来操作数据的方式是不可行的。例如，前面介绍过的Hadoop，对于它存储的数据就无法直接通过SQL来进行查询。此外由于企业已经习惯了小数据的灵活处理方式，对转到Hadoop上的大数据处理一下子变得无所适从。

很多传统的数据库和数据仓库厂商（如Teradata、Qracle和MySQL等）正在研究解决办法，它们的思路是Hadoop将MapReduce的结果存储到关系型数据库中，然后再查询RDBMS，但是这样的解决方案体现不出Hadoop的高效性。

为了让专业分析技术人员通过SQL语言来操作和分析大数据，大数据查询技术——SQL on Hadoop发展了起来。SQL on Hadoop是建立在Hadoop上的SQL查询技术，既能保证Hadoop的高性能，又结合了SQL的灵活性。SQL on Hadoop处于起步阶段，Hadoop解决方案对于SQL语言支持的广度和深度各不相同，技术实践方式也多种多样。最基本的方法是把传统的SQL语言通过中间转换后再进行操作。例如当前热门的Hive，它是一个数据仓库基础构架，建立在Hadoop上。另外，Hive还提供一系列的工具，从而实现了数据提取、转化、加载的功能，它可以存储、查询和分析存储在Hadoop中的大规模数据。Hive是把SQL语言（HiveQL，Hive的类SQL语句）编译成MapReduce，从而读取和操作Hadoop上的数据。SQL on Hadoop的基础技术提供了一种能力，让企业的信息管理能力从结构化数据拓展到非结构化数据。SQL on Hadoop技术仍在不断地发展，IT厂商也推出很多针对Hive的优化和扩展，它们大致分为以下3种情况。

一是Hive的性能改进和优化。Stinger、Presto也类似于Hive，将ANSI SQL编译成MapReduce。Stinger通过对原来的Hive做改进，优化了SQL查询的速度，使其完成对SQL查询时仅需要5～30秒。而Presto设计了一个简单的数据存储的抽象层，来满足在不同数据存储系

统（包括HBase、HDFS、Scribe等）上都可以使用SQL进行查询。

二是基于PostgreSQL的Hadoop分析。如Hadapt的技术（也称为DB on TOP），它为用户分析Hadoop里面的数据提供了一体化的分析环境，不仅可以分析SQL环境中传统的结构化数据，还可以解决非结构化数据。Hadapt是由Hadapt公司在2010年提出的，以EMC Greenplum HAWQ、Hadapt、Citus Data为代表。它结合了Hadoop环境和关系数据库环境，充分发挥了两者的优点，即Hadoop的可扩展性和关系数据库技术的高速性，它还可以在Hadoop层和关系数据库层之间自动划分查询执行任务。

三是实时交互SQL分析，如Apache的Impala和Drill。Apache的Drill项目最初的目标是建立共同的API和制定架构来容纳更多数据源、数据格式和查询语言，已经获得专业MapR的支持。Impala是Cloudera在受到Google的Dremel启发下开发的实时交互SQL大数据查询工具，它可以看成是Google Dremel架构和MPP（Massively Parallel Processing）结构的结合体。

5.1.1　Hive：基本的查询技术

Hive是由Facebook开发的建立在Hadoop上的数据仓库基础构架，是用来管理结构化数据的中间件。它架构在Hadoop之上，以Map-Reduce为执行环境，数据存储在HDFS上，元数据存储于RDBMS中。它提供了一系列用于存储、查询和分析大规模数据的工具。Hive适用于长时间的批处理查询分析。

Hive以SQL作为数据仓库的工具，具有很好的可扩展性、互操作性和容错性。可扩展性表现在可以自由扩展集群的规模，一般情况下不需要重启服务，而且支持用户自定义函数。用户可以根据自己的需求来实现自己的函数；互操作性表现在它是一个可以支持不同的文件和数据格式可扩展的框架；良好的容错性表现在当节点出现问题时，SQL仍可完成执行的操作。

Hive的本质其实是一个SQL解析引擎，它将SQL语句转译成MapReduce的工作，然后在Hadoop执行，来达到快速开发的目的。

1. Hive的体系结构

Hive主要分为4个部分，如图5.1所示[①]。

（1）用户接口

用户接口由三部分组成：Hive命令行接口CLI、Client和WebUI，其中最常用的是CLI。

① 命令行接口（CLI）：CLI启动的时候会同时启动一个Hive副本，CIL主要包括如下内容。

● DDL（数据描述语言）

生成表（creat table），放弃表（drop table），表改名（rename table）；

图5.1　Hive体系结构

① http://p-x1984.iteye.com/blog/761376

变更表（alter table），增加列（add column）；

增加分区（add Partitions）。

- Browsing（浏览）

查看所有表（show tables）；

查看表结构（describe table）；

查询（Queries）；

装载数据（Loading Data）。

② Client

Hive的客户端是Client，通过Client可以连接到Hive Server（服务端）。用户能够在启动Client模式的时候，找出Server所在的节点，同时还需要在该节点启动Server应用。用户可以通过JDBC/ODBC等驱动器来访问Hive，其中包括以下几项。

- Hive JDBC Driver（Java）
- Hive Add–in for Excel（by Microsoft）
- Hive ODBC Driver（C++）
- Thrift（C/C++、Python、Perl、PHP等）

③ Web UI

Web UI是指通过浏览器来访问Hive。主要包括以下两方面。

- MetaStore UI，它能导航和浏览系统中所有的表，并且可以给所有表和所有列做注释，也能抓取数据间的依赖关系。
- HiPal，它允许用户通过鼠标点击的方式交互地构建SQL查询，并支持投影、过滤、分组和合并功能。

（2）元数据存储（MetaStore）

- 存储表/分区的属性

Hive将元数据存储在RDBMS，如MySQL、Derby等，或者文本文件中。Hive中的元数据包括表、序列化和反序列化SerDe库，表的属性（是否为外部表等），表的名称，表的列和分区及其属性，表的数据所在HDFS的目录等信息。

- Thrift API

当前客户机的PHP（Web接口）、Java（查询引擎和CLI）、Python（旧的CLI）、Perl（Tests）等Thrift API。

（3）完成HiveQL查询的解释器、编译器、优化器、执行器

解释器（Parser）、编译器（Compiler）、优化器（Optimizer）是用于完成Hive QL查询语句，包括语法分析、词法分析、编译、优化和查询计划的生产。生产的查询计划主要存储于HDFS中，接着就会有MapReduce调用执行器（Executor）来执行。

（4）Hadoop

Hive的所有数据都存储于HDFS，大多数的查询由MapReduce来完成计算（该查询包括*的查询，比如select * from tbl就不会生成MapReduce任务）。

2. Hive的数据存储

Hive的数据存储在HDFS中，它没有指定的数据存储格式，也不需要像传统的SQL一样为数据建立索引，用户能极其自由地组织Hive中的表，仅仅需要在创建表的时候明确Hive数据中的行与列分隔符，Hive即可以解析数据。

Hive中包含以下四个数据模型：表（Table）、外部表（External Table）、分区（Partition）和存储桶（Bucket）。

（1）表（Table）

在概念理论上，Hive中的Table与数据库中的Table是相似的，所有的Table在Hive中都有着相对应的一个HDFS中的目录存储数据。比如，一个表tbs，在HDFS中它的路径是/wh/tbs，其中，wh是由${hive.metastore.warehouse.dir}指定的数据仓库的目录，除了External Table外，其他全部的Table数据都保存在该目录中。

Table包括创建过程以及数据加载过程（这两个过程都可以在同一个语句中完成）。在数据加载的过程中，实际数据会被移动至数据仓库目录中，而后面对数据的访问就会在数据仓库目录中直接完成。因此，在删除表时，表中所有的数据（包括元数据）都将同时会被删掉。

（2）分区（Partition）

Partition与数据库中的Partition列的密集索引相对应，然而在Hive中Partition的组织方式与在数据库中的有所不同。在Hive中，表中的一个Partition与表下的一个目录相对应，每个Partition的数据都有一个对应的目录。比如：tbs表中含有两个Partition: ds与city，那么对应于ds=20140801，city=US的HDFS子目录为：/wh/tbs/ds=20140801/city=US；而对应于ds=20140801，city=CA的HDFS子目录为；/wh/tbs/ds=20140801/city=CA。

（3）存储桶（Bucket）

Buckets用于计算指定列的hash值，然后按照hash值进行数据切分，达到并行的目的。每一个Bucket对应一个文件。比如，要将某个列分散为32个Bucket，首先对该列的值进行hash值计算，其中对应hash值是0的HDFS目录为/wh/tbs/ds=20140801/city=US/part-00000；hash值是20的HDFS目录为/wh/tbs/ds=20140801/city=US/part-00020。

（4）外部表（External Table）

External Table指向的是已经存在的HDFS中数据，而且可以创建分区和表。它和Table在元数据的组织上是一样的，但在实际数据的存储上却有不少的差异。

与Table相对比，External Table仅有一个过程，加载数据和创建表是同时进行和完成的（CREATE EXTERNAL TABLE ……LOCATION），而且实际数据存储在LOCATION后面指定的HDFS路径中，并不会移动至数据仓库目录中。当一个External Table被删除时，仅仅删除元数据，表中的数据不会真正被删除。

3. Hive QL

Hive定义了简单的类SQL查询语言，称为Hive QL，也缩写为HQL。由于SQL在数据仓库中应用的很广泛，因此，开发者专门针对Hive的特性设计了类SQL的查询语言HQL。这样方便了熟悉SQL开发和MapReduce的开发者使用Hive进行开发和自定义来处理复杂的分析工作，而且还可以利用HQL进行用户查询。

下面简单介绍Hive QL的一般查询语句的基本格式。

```
   SELECT [ALL | DISTINCT] select_expr,  select_expr,  …
 FROM table_reference
[WHERE where_condition]
 [GROUP BY col_list]
[
 CLUSTER BY col_list  | [DISTRIBUTE BY col_list]
    [SORT BY col_list]
]
[LIMIT number]
```

其中，SELECT语句可以是union查询或子查询的一部分。选用ALL或者DISTINCT选项区分的目的是对重复记录进行处理，默认情况是ALL，代表查询所有记录；DISTINCT表示去掉重复的记录。table_reference是查询的输入，输入方式可以是一个普通table、一个视图、一个join或子查询。where condition是指一个布尔表达式，但是Hive不支持在WHERE子句中的IN，EXIST或子查询。Limit用于限制查询的记录数。查询结果是随机选择的。

注：Hive的官方文档对查询语言有很详细的描述，具体请参考http://wiki.apache.org/hadoop/Hive/LanguageManual

Hive QL常用的查询操作主要包括ANSI JOIN（只有equi-join）、多个表Insert、多个表的Group by和Sampling等。其中，Join和Group by操作的说明如下。

（1）Join操作

Join操作的示意图如图5.2所示。

Page_view

pageid	User·id	time
1	11	10:01
1	12	10:08
2	11	10:01

×

user

User·id	age	gender
11	18	female
12	24	male

=

pv_users

pageid	age
1	18
1	24
2	18

INSERT INTO TABLE pv users
SELECT pv.pagied，U.age
FROM Page_view pv JOIN User U ON（pv.user id=U.user id）；

图5.2 Join操作

例如，通过图5.2中一个简单的语句表示一个3列的表Page_view（访问网页id、用户ID和访问时间）和3列的表User（用户ID、年龄和性别），通过相同的用户ID执行Join的操作，形成一个新的表pv_users，该表展示了访问页面的用户年龄结构。

在使用Join操作的查询语句时要注意：应将条目少的表或者子查询放在Join操作符的左边。由于在Join操作的Reduce阶段，Join操作符左边的表里面的内容会被加载进内存，因此，将条目或者子查询少的表放在左边，这样可以有效减少发生OOM错误的几率。

（2）Group by操作

Group by的操作如图5.3所示。

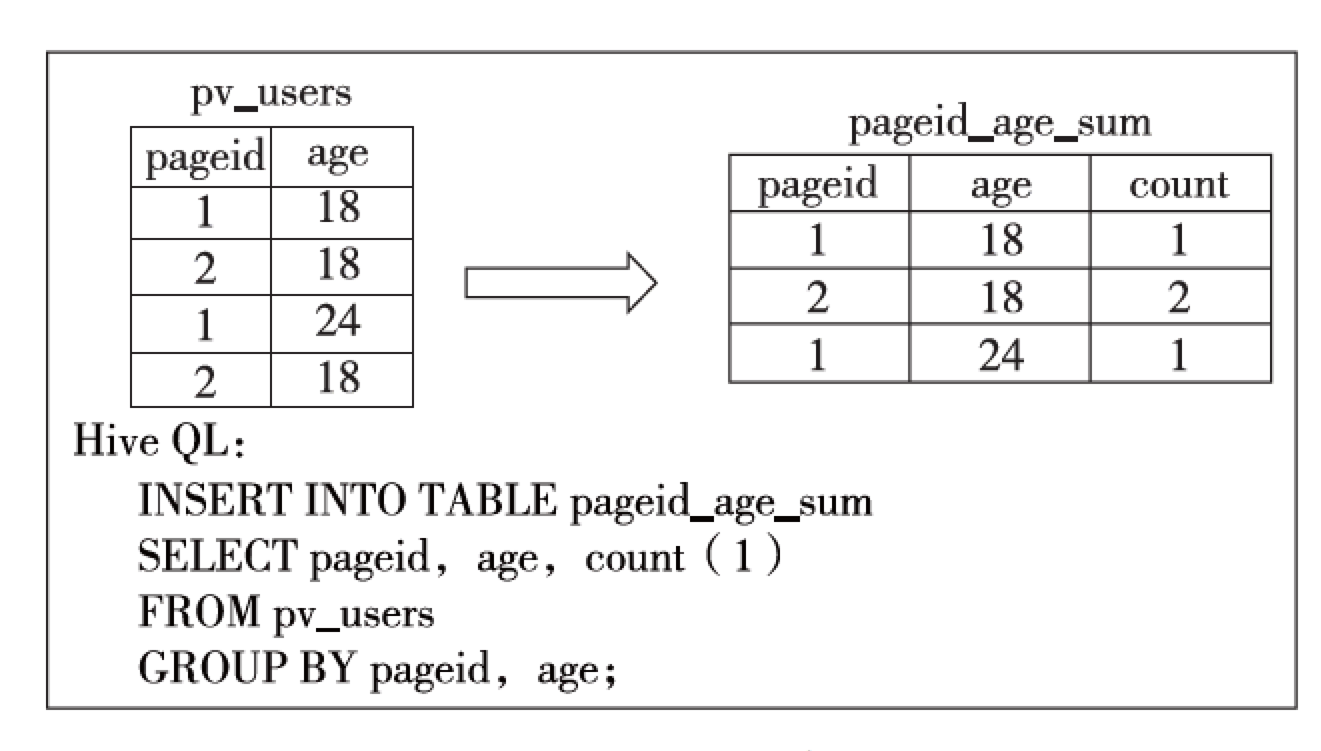


图5.3　Group by操作

5.1.2　Hive的优化和升级

Facebook在2007年提出Apache Hive和Hive QL后，它们就成为事实上的Hadoop上的SQL接口。如今，各种类型的公司都在使用Hive来访问Hadoop数据，希望给公司或者用户带来更多的价值。此外，还有许多的公司利用大量的BI工具来达到同样的目的，这些BI工具也是使用Hive作为接口。因此，Hive主要用于建立大规模的批计算，这在数据报告、数据挖掘以及数据准备等应用场合十分有效。但是随着Hadoop的需求越来越广阔，企业用户也越来越需要Hadoop具备更高的实时性和交互性。

目前Hive还存在很多不足，特别是在查询速度方面有着“先天性不足”。在查询过程中，面对一个完整的数据集可能要花费几分钟到几个小时，这是完全不切实际的。对于主流用户而言，也很难有大的吸引力。

Hive的优化和升级是很多IT企业的一项重要工作，其中主要的项目包括Stinger Initiative和Presto等。

1. Stinger

Stinger是Hortonworks开源的一个类SQL的即时查询系统，是对Hive进行优化的项目。Hortonworks声称较Hive可以提升100倍的速度。它主要的改进如下。

（1）库优化：智能优化器

- 生成简化的有向无环图。
- 引入in-memory-hash-join，适用于有一方适合在内存中的Join。这是一个全新的in-memory-hash-join算法，借此算法Hive可以把小表读到哈希表中，可以遍历大文件来产生输出。
- 引入sort-merge-bucket-join，适用于表在同样的关键词上被分为bucket的情形，在速度改进方面是巨大的。
- 减少在内存中的事实表的足迹。
- 让优化器自动挑选map joins。

（2）多维度的结构化数据

在Hive中采用企业级数据仓库（EDW）中很普通的维度模式，产生大的数据表和小的维度表。维度表经常小到能适合RAM，有时被称为Star Schema。

（3）优化的列存储（ORCFile）

优化的列存储包括如下内容。

- 生成一个更好的列存储文件，与Hive数据模型紧密一致。
- 把复杂的行类型分解为原始类型，便于更好地压缩和投影。
- 对于必需的列，从HDFS中只读bytes。
- 既储存文件也储存文件的每个节。
- 增加了聚合函数，如min、max、sum、average和count等。
- 允许通过排序列快速访问，能够快速校验一个值是否存在。

（4）深度分析能力

支持SQL：2003 Window Functions。

其OVER子句支持Multiple PARTITION BY 和ORDER BY；支持Windowing（ROWS PRECEDING/FOLLOWING）；支持大量的聚合，如RANK、FIRST_VALUE、LAST_VALUE、LEAD/LAG、Distrubutions等。

（5）与Hive数据类型的一致性

数据类型的优化包括如下内容。

- 增加了固点NUMERIC和DECIMAL类型。
- 增加了有限域大小的VARCHAR和CHAR类型。
- 增加了DATETIME。
- 对FLOAT增加大小从1～53。
- 增加了考虑兼容性的同义字，对应BINARY的BLOB，对应STRING的TEXT，对应FLOAT的REAL。
- 增加了SQL语义，如更多地用IN、NOT IN、HAVING的子查询，EXISTS和NOT EXISTS等。

（6）架构的优化

Stinger架构与Hive不同的是，Stinger采用Tez。所以，Hive是SQL on Map-Reduce，而Stinger是Hive on Tez。Tez的一个重要作用是优化Hive和PIG这种典型的DAG应用场景，它通过减少数据读写I/O操作，优化了DAG流程，使得Hive的速度提高了很多倍。其架构如图5.4所示。Stinger是在Hive的现有基础上加了一个优化层Tez（此框架基于Yarn），所有的查询和统计都要经过它的优化层来处理，以减少不必要的工作以及资源开销。虽然Stinger也对Hive进行了较多的优化与加强，Stinger的总体性能还是依赖于其子系统Tez的表现。而Tez是Hortonworks开源的一个DAG计算框架，Tez可以理解为Google Pregel的开源实现，该框架可以像Map-Reduce一样，用来设计DAG应用程序，

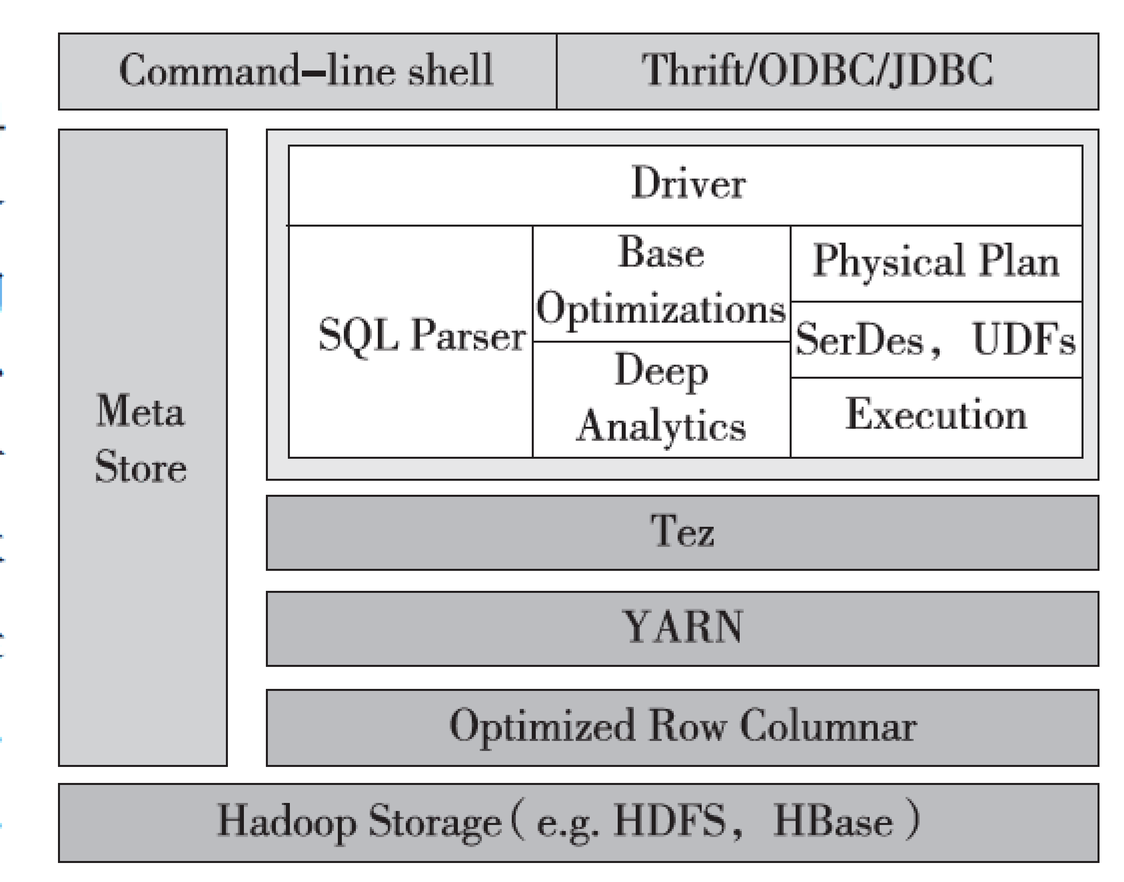


图5.4　Stinger架构

但需要注意的是，Tez只能运行在YARN上。

2. Presto

2013年11月Facebook开源了一个分布式SQL查询引擎Presto，它专门用来进行快速、实时的数据分析。Presto支持标准的ANSI SQL子集，包括复杂查询、聚合、连接和窗口函数。其简化的架构如图5.5所示，客户端将SQL查询发送到Presto的协调器，协调器会对SQL查询进行语法检查、分析和规划查询计划。调度器将执行的管道组合在一起，将任务分配给那些离数据最近的节点，然后监控执行过程。客户端从输出段中将数据取出，这些数据来源于更底层的处理段中。因此，使用Presto进行简单查询仅仅需要几百毫秒，即使是复杂的查询，也只需要几分钟就可以完成。

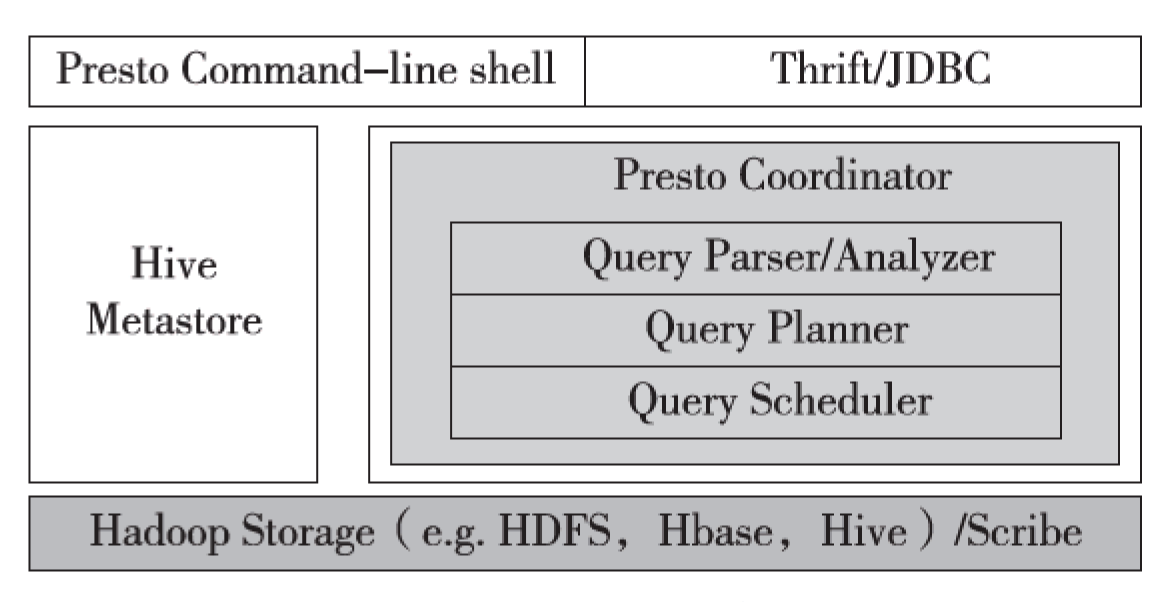


图5.5　Presto架构

Presto的运行模型与Hive有着本质的区别。Hive是将查询转换成多阶段的Map–Reduce任务，依次运行。每一个任务在磁盘上读取输入数据并且将中间结果输出到磁盘上。然而Presto引擎没有使用Map–Reduce，它使用了一个定制的查询执行引擎和响应操作符来支持SQL语法。除了改进的调度算法之外，所有的数据处理都是在内存中进行的。不同的处理端通过网络组成处理的流水线。这样可避免不必要的磁盘读写和额外的延迟。这种流水线式的执行模型会在同一时间运行多个数据处理段，一旦数据可用的时候就会将数据从一个处理段传入到下一个处理段。这种方式会大大减少各种查询的端到端的响应时间。此外，Presto设计了一个简单的数据存储抽象层，来满足在不同数据存储系统上都可以用SQL进行查询的需要。存储连接器目前除支持Hive/HDFS外，还支持HBase、Scribe和定制开发的系统。

但是在功能上，Presto与Hive有点差别，也可以说是Presto功能较不完善，毕竟Presto推出时间不长，这些差别可以概括为：

（1）Presto完全没有数据写入功能，不能使用create语句创建表，（可通过CREATE TABLE tablename AS query）建立视图、导入数据。

（2）Presto不支持UDF（用户自定义函数）。

（3）Presto支持窗口函数，但数量与Hive比相对较少。

5.1.3　实时交互式SQL查询

数据库进行实时交互SQL查询。

1. Impala

Impala是Cloudera在受到Google的Dremel启发下开发的实时交互式SQL大数据查询工具，根据其官网的产品实测表明Impala的查询速度比原来的Hive QL提升3 ~ 90倍。Impala属于交互式的SQL查询，比SQL的查询功能更胜一筹，它可以看成是Google Dremel架构和MPP (Massively Parallel Processing)结构的结合体，主要是为了在Hadoop上进行低时延的SQL查询而专门设计

的。它不仅采用接近ANSI-92标准的Hive QL来执行查询，而且具有对Hadoop应用兼容的SQL接口。Impala通过使用与商用并行关系数据库中类似的分布式查询引擎，代替了使用较缓慢的Hive+MapReduce的批处理方式。

Impala架构如图5.6所示。Impala采用的接口主要是Hive QL和JDBC/ODBC等，具有全局统一的元数据存储和调度，查询则是完全的分布式并行处理，对HDFS或者HBase直接在本地读取。Impala分布式查询引擎主要包括Query Planner、Query Coordinator和Query Exec Engine三部分，Query Planner接收来自SQL APP和ODBC的查询，然后将查询转换为若干个子查询；Query Coordinator将这些子查询分发到各个节点上；由各个节点上的Query Exec Engine负责子查询的执行，最后返回子查询的结果，这些中间结果经过聚集之后最终返回给用户。

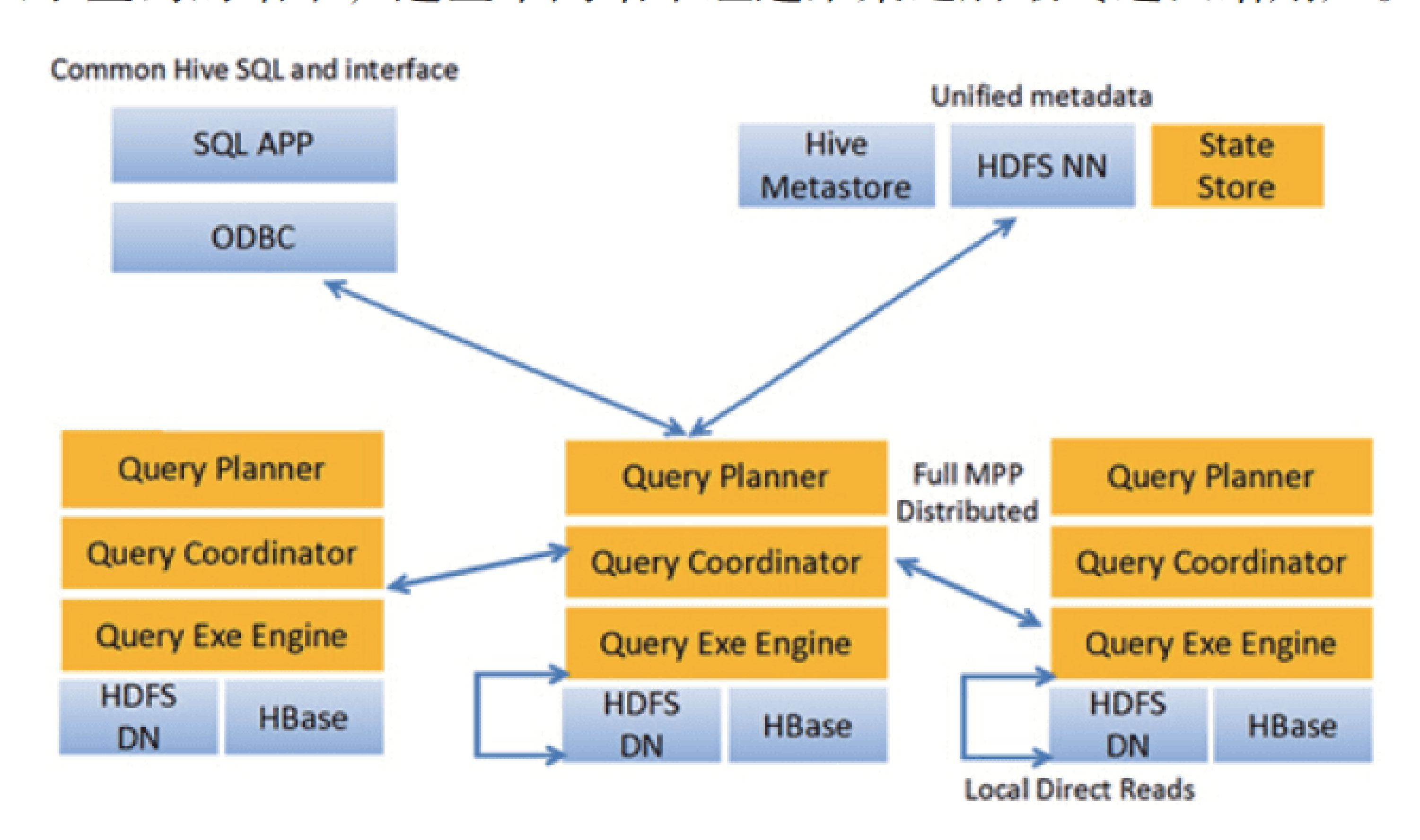


图5.6　Impala架构

Impala可以从HDFS或HBase中通过Select、Join和统计函数等方法查询数据，从而大大提高了效率。由于直接在HDFS或者HBase存储和元数据上做查询，使得它具有Hadoop的灵活性、横向扩充性和低成本优势，而且不需要数据和元数据的复制或同步，这样的本地处理有效地避免了网络瓶颈。

Impala主要由Impalad、State Store和CLI三部分组成。Impalad与DataNode运行在同一节点上，Impalad是每个节点上处理数据的基本单元，由Impalad进程表示，它接收客户端的查询请求（接收查询请求的Impalad为Coordinator，Coordinator通过JNI调用Java前端解释SQL查询语句，生成查询计划树，再通过调度器把执行计划分发给具有相应数据的其他Impalad来执行），读写数据，并行地执行查询，之后把结果通过网络的流式传送回给Coordinator，由Coordinator返回给客户端。同时Impalad也与State Store一直保持连接，用于确定哪个Impalad是健康的并且可以接受新的工作。

Impala的另一个核心组件State Store负责检测整个集群中所有节点上的进程的健康度。State Store通过连续不断的分发findings到每一个节点上的进程。State Store的物理进程名称为statestored，在一个Impala集群中仅需要一个这样的进程。如果Impala集群中有一个节点因为硬件故障、网络错误、软件问题或是别的原因导致该节点不可用，则State Store通知所有在集群中的其他正常的节点，以便在新任务提交的时候可以避免将新任务分发到该故障节点上。由于State Store的应用场合是在集群发生故障的时候通知集群中其他的节点，来避免在新的任

务到来时把任务发送到故障节点上，因此State Store不是关键的操作。如果State Store没有运行或是连接不上，其他的节点仍可以继续运行分布式的分发和处理任务，结果就是集群的鲁棒性上受到一些影响。当State Store恢复的时候会继续和其他的节点通信然后恢复其监控函数。

CLI提供给用户查询使用的命令行工具，同时Impala还提供了Hue、JDBC、ODBC和Thrift使用接口。

2. Drill

大数据查询并不是单一的交互式查询，大多数查询都是很缓慢而且非交互式的。为了解决这一问题谷歌研发了Dremel，它能以极快的速度处理大规模的海量数据。根据谷歌的研究报告显示，Dremel只需几秒就可以完成PB数量级的查询，而Drill就是其对应的Apache开源版本。Drill是一个专门用于互动分析大型数据集的分布式系统，它采用标准的SQL语句来进行大数据的互动分析，客户端可通过任何一个SQL工具输入SQL语句；通过Drill的ODBC驱动器连接到Drill的驱动，并经过SQL查询的Parser进行解析；经过Query Planner进行计划，在节点的Drilbit上来执行。多个Drilbit共同完成大规模的查询任务，体现出Drill具有查询快、开放性和现代化等特点。

5.1.4　基于PostgreSQL的SQL on Hadoop

Hadapt、EMC Greenplum HAWQ和Citus Data都是基于PostgreSQL的解决方案。其中Hadapt是专注于SQL on Hadoop的厂商。Hadapt是一个自适应分析平台，具有健壮性、可伸缩性等特点，为Apache Hadoop开源项目带来了SQL实现。将数据在两种计算框架中分别存放是Hadapt解决方案的本质，如图5.7所示。结构化数据存储于高性能关系型数据库中，非结构化数据存储于Hadoop分布文件系统，对两种类型的数据交互依靠查询的切片来执行，通过它合并了关联数据存储的混合存储层。

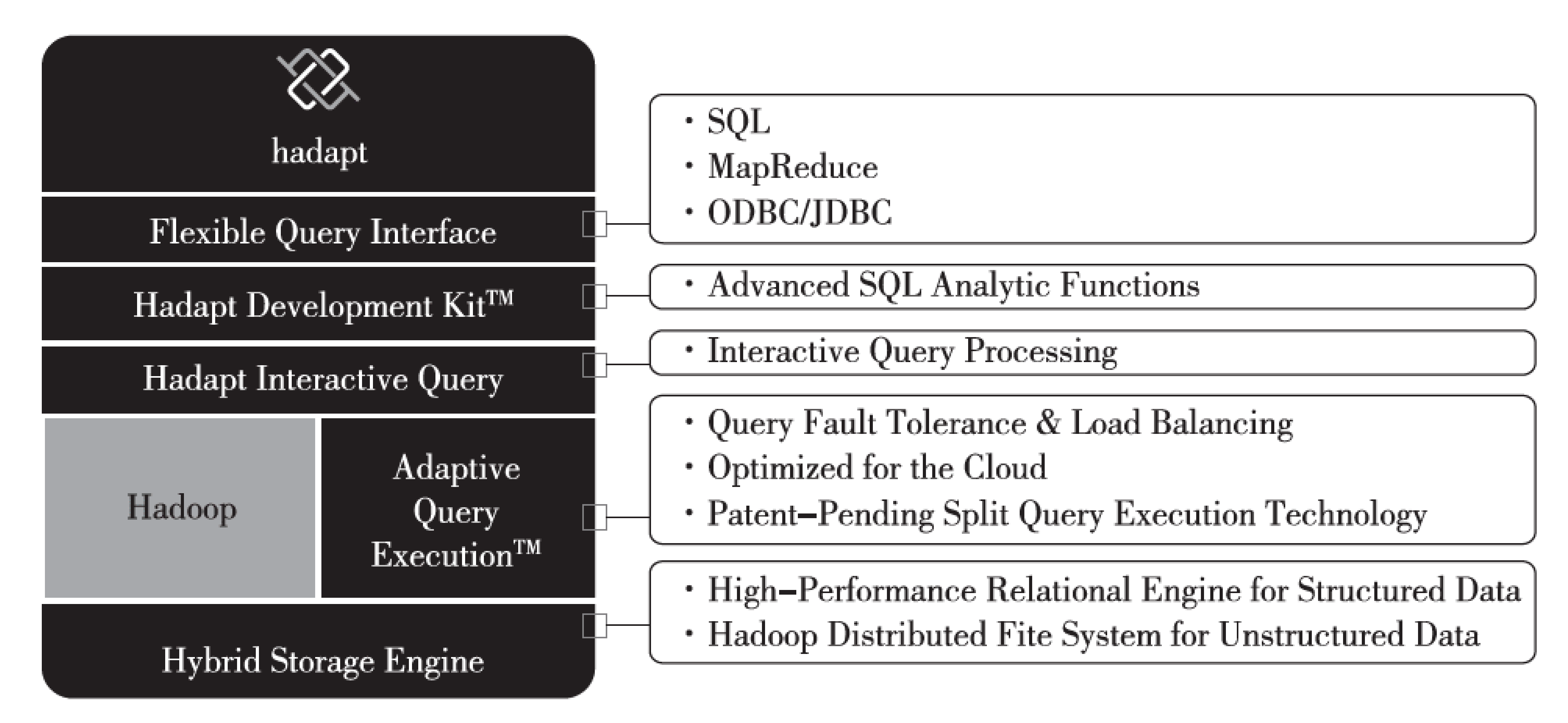


图5.7　Hadapt的体系架构

Hadapt允许进行基于SQL大数据集的交互分析。通过Hadapt交互式的查询，可以自定义分析的Hadapt Development Kit™（HDK）和Tableau软件集成，Hadapt 2.0成为Hadoop工业上第一个交互式应用程序。

Hadapt结合了Hadoop和关系数据库管理软件的优点，成为一个单独的数据平台。Hadapt统一了关系数据库环境和Hadoop环境，可以在Hadoop层和关系数据库层之间自动划分查询执行任务，使得它具有一体化的分析环境，这样不仅可以对Hadoop里的数据执行分析操作，还能对SQL环境中传统的结构化数据进行分析，并充分利用了Hadoop的可扩展性和关系数据库技术的高速度。Hadapt公司表示，Hadapt通常采用的方法是使用由扩充型连接件联系起来的两个不同系统，但是这带来了延迟，导致这种方法显得很孤立。而Hadapt的平台设计成了可以在私有云或公共云环境上运行，具有在同一个环境中就能访问所有数据的优点。在Hadapt中除了MapReduce流程和大数据分析工具外，现有的基于SQL的工具也可以使用。

5.2 数据分析的方法与技术

20世纪初期，进行数据分析是一件非常困难的事情。如果要对某一问题进行深入的分析，例如，建立模型，实现预测功能，则完全需要依靠人们手工进行各种统计运算。因此，那个时代几乎没有人拥有可扩展的数据分析能力。数十年过去了，现在人们处理数据的规模已经远远超过手工处理时代的数据规模，因此，手工进行数据分析计算已经完全无法满足时代的需求了。计算机技术的发展有效地解决了这一难题，随着数据规模的迅速扩大，计算机处理数据的能力也在不断增强，人们已经从手工时代进入为半自动化（有的已经达到自动化）时代了。目前，社交网站、电子商务等网络服务的快速发展，使得网络服务与网络信息规模呈裂变式增长，这样大规模的数据也给计算机与数据分析带来了巨大的挑战。面对这个巨大的挑战，很多IT厂商在不断地找寻解决方案，人们把这个时代定义为大数据时代。

众所周知，大数据不单是数据量大的事情，最重要的是怎么利用好这些大数据，也就是对大数据进行分析，只有通过分析这些数据才能获取更多智能的、深入的、有价值的信息。现在越来越多的行业应用涉及到了大数据，例如金融、零售业、医疗、电信、航空等。这些行业应用不断地产生大量数据，而这些数据的属性，包括数量，速度，多样性、复杂性等都在呈现不断增长的复杂性，所以大数据的分析方法在大数据领域就显得尤为重要，可以说是决定最终信息是否有价值的决定性因素。基于此，数据分析普遍存在的方法理论有哪些呢？

数据分析是指采用准确适宜的分析方法和工具来分析经过处理的数据，提取有价值的信息，从而形成有效的结论并通过可视化技术展现出来的过程。因此，要学习数据分析必须首先需清楚数据分析与数据展现的方法以及对数据分析工具的使用。数据分析的方法大致可以分为三种：基本分析方法，该类方法主要以基础的统计分析为主；高级分析方法，以计量经济建模理论为主；数据挖掘类，以数据仓库、机器学习等复合技术为主[①]。数据仓库技术前面已经讲到，数据分析最重要的领域——数据挖掘技术在后面章节会着重介绍，本节主要介绍基本分析方法和高级分析方法。数据展现主要是指通过数据可视化技术把数据转换成图形或图像在屏幕上显示出来。现在普遍使用的数据分析工具包括：Excel、SPSS、SAS、Eviews、R语言、MATLAB、Stata和Weka等，具体的应用工具软件会在5.3节中详细介绍。本节主要介绍数据分析的方法以及可视化技术。

① http://wenku.baidu.com/view/e42e03679b6648d7c0c74604.html.

5.2.1　基本分析方法

目前，数据分析方法中常见的基本分析方法包括对比分析、趋势分析、差异显著性检验、分组分析法、结构分析、因素分析法、交叉分析法、综合评价分析、漏斗图分析法等。

1. 对比分析

对比分析也称为比较分析，该方法通过对客观事物进行对比，从而认识事物的本质以及挖掘事物的规律并且给出准确的评价。对比分析的分析对象一般为相互联系的两个指标数据，它主要展示与说明研究对象水平的高低、速度的快慢、规模的大小以及各关系之间是否协调。

对比分析的分类分为：横向对比、纵向对比、标准对比以及实际与计划对比。横向对比是指同一时间条件下不同总体指标比较，如不同公司、不同地区、不同国家的比较，也叫静态对比。纵向对比是指同一总体条件不同时间指标数值的比较，也叫动态对比。这两种方法既可单独使用，也可结合使用。进行对比分析时，可以使用的指标包括总量指标、相对指标、平均指标，这些指标可以单独使用也可结合起来使用。分析比较的结果可用相对数，如百分数、倍数、系数等来反映；也可用相差的绝对数和相关的百分比来表示，即将对比的指标相减。标准对比是指实际指标与标准水平进行对比，了解当前的指标和标准指标的差异，此处标准水平是根据由经验或理论得出来的。实际与计划对比是反映实际与目标值的差异，主要是利用当前实际值与目标的计划数、预算数、指标数等对比从而得到差异。

对比分析在了解财政收支数据特征方面很有优势，同时该方法还可以用于差异分析，在对比时，不仅可以使用单指标，还可以使用多指标进行综合评价[①]。值得注意的是，合适的对比标准是成功的关键，因此在选择对比的对象时需要根据分析的目的选择合适的对比标准，使得指标上具有可比性。

2. 趋势分析

趋势分析是指将实际达到的结果，通过比较同类指标不同时期的数据，继而明确该指标的变化趋势以及变化规律的一种分析方法。趋势分析主要是运用在财务分析方面，具体的分析方法包括定比和环比两种方法。

定比分析是报告分析期的水平比上某一特定时期的水平，它阐释的是该现象在不短的一段时期内总的变化水平[②]。在实际工作中，该方法主要用于分析年度发展变化的速度情况。其中分析的重要的指标是定基动态比率，该计算公式为：定基动态比率＝分析期数值÷固定基期数值。

环比分析指的是报告分析期水平比上前一时段水平，表示是逐期变化趋势的现象，然后通过本期数据与上期数据的对比，形成时间序列图。环比分析能够反映逐期的变化情况，但会受很多因素的影响，其中最主要的是季节影响，这时就会出现大幅度波动，不能真实反映变化趋势。因此，环比分析适用于没有季节因素的时间序列数据。环比分析中重要的指标是

① http://wenku.baidu.com/view/90f012d176a20029bd642d71.html.

② http://wenku.baidu.com/view/e42e03679b6648d7c0c74604.html.

环比动态比率，该比率是通过分析期和它的前期数值相除计算出来的比率，其计算公式是：环比动态比率＝分析期数值÷前期数值。

3. 显著性检验

差异显著性检验（Significance Test）是指事先对总体（随机变量）的参数或者总体分布形式做一个假设，然后利用样本信息来判断该假设（原假设）是否合理，即判断总体的原假设与真实情况是否存在显著性差异。或者说，显著性检验是判断样本与对总体所做的假设之间的差异是属于机会变异，还是由所做的假设与总体真实情况之间不一致而产生的差异。因此，显著性检验是对总体所做的假设是否合理正确所进行的检验，其原理是用“小概率事件实际不可能性原理”来接受或否定假设。

常用的检验方法包括：t检验、方差分析、μ检验、零反应检验等，本章主要介绍t检验和方差分析。

（1）t检验

t检验是通过t分布理论来推断差异发生的概率，继而比较两个平均数之间是否存在显著差异。它与z检验、卡方检验并列。t检验的类型主要包括：单样本t检验、独立样本t检验、配对样本t检验。

①单个样本的t检验

单个样本检验是检验一个样本平均数和已知的总体平均数是否存在显著差异。

目的：比较代表未知总体平均数的样本平均数μ和已知总体均数μ_0。

计算公式如下。

t统计量：$t=\dfrac{\mu-\mu_0}{s/\sqrt{n}}$；自由度：$v=n-1$ （5–1）

说明：μ为样本平均数；μ_0为总体平均数；s为样本标准差；n为样本容量。

适用条件：一个总体均数是已知的，能够得到一个样本均数和该样本标准差，样本来自正态分布或近似正态分布总体。

②配对样本t检验

配对样本t检验用于检验两个相关样本或成对样本所得均值间的差异是否有统计搜索学意义，但在检验之前必须先对样本进行配对设计，将受试对象的某些重要特征按照相近的原则配成对，其目的是为了消除混杂因素的影响。观察对象之间除了处理因素和研究因素之外，其他因素基本相同，每对中的两个个体随机给予两种处理。配对样本t检验包括两种情形，一是同源配对，即同一受试对象或同一样本分为两个部分，分别接受两种不同的处理后进行对比，目的是判断不同的处理是否有差别；二是自身配对，某种同质对象分别接受两种不同的处理，或者某种同质对象接受处理前后是否有差异，即对同一受试对象在处理前后的结果进行比较。

计算公式：$t=\dfrac{\bar{d}}{s_d/\sqrt{n}}$

自由度：$v=n-1$ （5–2）

说明：$\bar{d}$，差值的样本平均数；s_d，差值的标准差；n为对子数。

适用条件：两个样本是配对的，而且样本所在的总体符合正态分布。

（2）方差分析

方差分析（Analysis of Variance，简称ANOVA），又称“F检验”或“变异数分析”，是通过比较总体方差各种估计间的差异，来检验方差的正态总体是否有相同的均值。方差分析是用于两个或两个以上样本均数差别的显著性检验，其本质主要是通过分析研究变量之间的关系，不同来源的变异对总体变异的贡献大小，从而确定可控因素对研究结果影响力的大小。方差分析主要包括单因素方差分析和多因素方差分析。

方差分析的基本步骤如下：

（1）建立检验假设；H_0：多个样本总体均数相等；
H_1：多个样本总体均数不相等或不全等；
其检验标准为0.05；

（2）计算检验统计量F值；

（3）确定P值并作出推断结果。

方差分析是研究分类性自变量对数值型因变量的影响，通过显著性来确定变量之间是否有关系，以及关系强度如何。方差分析广泛应用于农业、商业、医学、经济学等方面。

4. 分组分析

分组分析法是指通过统计分组的计算和分析，来认识所要分析对象的不同特征，不同性质以及相互关系的方法。

分组的方法主要是根据研究的目的和客观现象的内在特点，按照某个标志或几个标志把被研究的总体划分为若干个不同性质的组，使组内的差异尽可能小，组间的差异尽可能大。分组分析法是在分组的基础上，从定性或定量的角度对现象的内部结构或现象之间的依存关系做进一步的分析研究，以便寻找事物发展的规律，然后正确地分析问题和解决问题。

分组时需要遵循两个原则，即穷尽原则和互斥原则。所谓穷尽原则，就是使总体中的每一个单位都要被分到一个组中，也就是说各分组的空间要能够容纳总体所有的单位。互斥原则，是在特定的分组标志下，总体中的任何一个单位只能归属于某一个组，而不能同时归属于几个组。

5. 结构分析

结构分析是建立在对比分析的基础上，扩大对比范围，然后运用结构分析进行一一比较，通过结构指标来解释企业资源结构分布、生产布局的状况，便于经营者进行调整，投资者长期决策。

$$结构指标（\%）=（总体中某一部分 \div 总体总量）\times 100\%$$

结构指标是指总体某一部分占总体总量的比重，总体中各个部分的结构相对数的和等于100%①。

结构分析广泛应用于财政收支领域，它能够从不同的维度展开结构分析，如科目结构、区域结构等。同时饼图、圆锥图和金字塔图等都是开展结构分析的有效工具。

根据关注的时间，可分成静态结构分析和动态结构分析；根据关注的对象，结构分析可

① http://www.baike.com/wiki/结构性分析.

分成增量结构分析、元素的比重分析以及总量结构分析。结构分析还能够在多种分类间进行交叉结构分析。另外，按照结构内元素，结构分析可以是结构差异分析，并且可以自定义结构①，如图5.8所示。

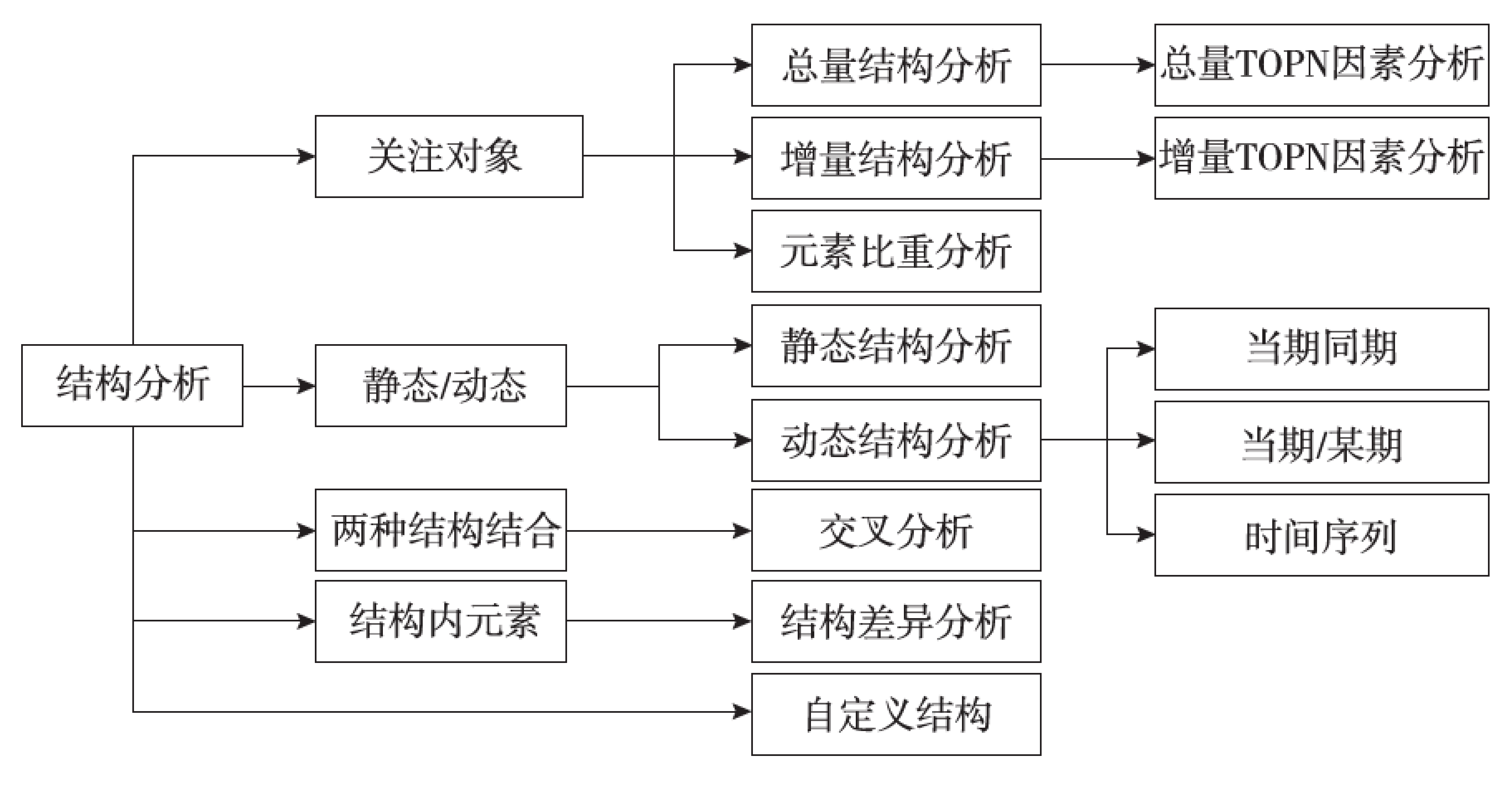


图5.8　结构分析分类图

6. 因素分析

因素分析法是斯皮尔曼（C.Spearman）在1904年提出的，根据分析指标与其影响因素的关系，从而确定不同因素对分析指标影响程度以及影响方向的一种方法。因素分析法既能够独立分析某个因素对经济指标的影响，又能够全面分析各因素对某一经济指标的影响，因素分析法实际就是相关性概念，是在心理学领域中发展起来的一种多变量解析手段。

因素分析有以下三种常见的方法。

（1）连环替代法，是把分析指标拆分为多个能够计量的因素，同时按照因素相互间的依存关系，依次测定各因素的比较值（一般即实际值）替换基准值（一般为计划值或者标准值），据以测定不同因素对分析指标产生的影响。

（2）差额分析法，又称绝对分析法，是基于连环替代法的一种简化形式，主要依据每个因素的实际值与标准值之间的差异，来计算不同因素对分析指标的影响以及影响的程度。

（3）定基替代法，分别用基期标准值替换实际值，就不同因素对指标产生的影响进行测定。与连环替代法不同的是，其替代的顺序恰好相反。

7. 交叉分析法

交叉分析法是指将有一定联系的两个变量及其值交叉排列在一张表内，使各变量值成为不同变量的交叉结点，形成交叉表，从而分析交叉表中变量之间的关系，也叫交叉表分析法。它是从交叉、立体的角度出发，由浅入深、由低级到高级的一种分析方法。虽然复杂，但这种方法弥补了“各自为政”分析方法所带来的偏差。常用的是二维交叉表分析法，当然也有二维以上的交叉表，维度越多，交叉表越复杂，所以在选择维度的时候需要根据分析目

① http://wenku.baidu.com/view/e42e03679b6648d7c0c74604.html.

的来决定。

目前交叉表分析法被广泛用于商业市场的调研工作，因为它有如下优点。

- 交叉表的分析结果容易直观地被理解。
- 明确的解释加强了调研结果与调研对象行为的联系。
- 一系列交叉表比多变量分析更容易理解复杂的问题。
- 交叉表可以减弱空格问题，这在多元离散变量分析中更突出。
- 交叉表可将复杂的数据简单化。

交叉表分析法也存在着局限性。第一，如果需要考虑多个变量，样本容量就应该相当大。第二，很难确保是否对所有的相关变量都进行了分析，如果变量选择的不适当，就会导致得出错误的结论。即使选择了正确的变量，研究者也会因使用不当而无法发现真正的关系。研究者选择关键变量以及根据这些变量组成交叉表的能力决定了其能否制作出一个好的交叉表。另外，随着研究的目的、性质的变化，用于交叉表分析的变量的类型和数量也应该变化。因此，交叉表分析只能用于有数据基础的变量分析，它描述的是变量间的关系，但不一定是因果关系。

8. 综合评价分析法

综合评价分析是运用多个指标对多个参评对象进行评价的方法，也称为多变数综合评价方法。其基本思想是将多个指标转化为一个能反映综合情况的指标来进行评价。如不同国家经济实力，不同地区社会发展水平，小康生活水平达标进程，企业经济效益评价等，都可以应用这种方法。它的特点包括以下四个方面。

- 评价过程不是逐个按照顺序完成，而是同时完成多个指标的评价。
- 在过程中，根据指标的重要性进行加权处理。
- 评价结果不是含有具体意义的统计指标，而是以指数表示参数。
- 评价对象综合状况的排序。

9. 漏斗图分析法

漏斗图分析法是一个适用于业务流程比较规范、周期比较长、各环节流程比较复杂、业务比较多的分析方法。在业务流程中使用漏斗图可以很快地发现业务流程中哪些环节存在问题，并且用一种直观的方式说明了问题的所在。

下面举例说明漏斗图用于某网站中某些关键路径的转化率分析，它不仅能够显示用户从进入网站到实现购买的最终转化率，同时还能够展现整个关键路径中每一步的转化率，如图5.9所示。

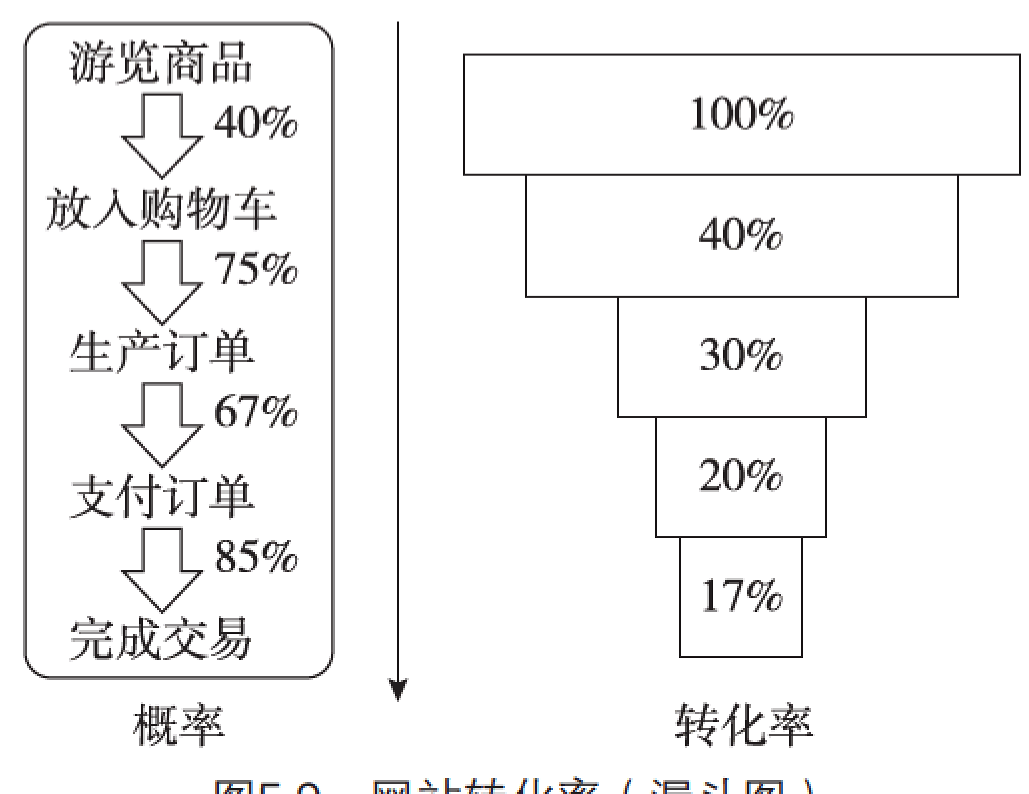


图5.9　网站转化率（漏斗图）

单一的漏斗图不能作为评价网站某个关键流程中各个步骤转化率的好与坏的方法。它可以结合前面讲到的对比分析法，对同一环节前后效果进行对比分析，或者对同一环节不同细

分用户群的转化率进行对比，也可以对同一行业类似产品的转化率进行比较等。

漏斗图不仅能体现用户在业务流程中的转化率和流失率，还可以告知相关人员哪种业务在网站中是最重要、最受欢迎的。通过对不同业务漏斗图的对比，同时灵活结合运用漏斗图与其他分析法，从而找出业务受欢迎程度的不同，发现隐藏在其中的问题。

5.2.2 高级分析方法

1. 时间序列分析

时间序列分析是一种对动态数据进行处理的分析方法，指一个依时间顺序组成的观察数据集合。它包括一般统计分析（例如谱分析、自相关分析等），建立并推断统计模型，以及有关时间序列的最优预测、控制和滤波等内容。比如医院每天门诊接诊人数序列、电信每日流量产生量序列、逐年人口统计资料等。传统的统计分析均假设数据序列具有独立性，然而时间序列分析更偏重于对数据序列的相互依赖关系进行研究。

一个时间序列一般由4种要素组成：趋势、季节变动、循环波动和不规则波动。趋势是指持续向上或者向下的波动，季节变动是指一年时间内周期性的波动，循环波动是指非固定长度的变动，不规则波动是指产生一种波浪形或震荡式的变动。时间序列主要的用途包括：系统描述、系统分析、预测未来、决策控制，其中预测是时间序列中最重要的内容。时间序列分析对数据资料要求很严格，不允许出现缺失值，所以时间序列分析还需要运用到缺失值填补的方法。

（1）时间序列的方法分类

时间序列分析方法通常分为描述性时序分析以及统计时序分析。

描述性时序分析是根据较为直观的数据或者绘图观测，从中找出序列所蕴含的发展规律。该方法比较简单，通常是操作统计时序分析的第一步。例如，通过对比公司财政收入的历年增长趋势和季节波动趋势（如图5.10、图5.11所示），可以发现同一个公司年度财政收入数据呈现持续稳定地增长，季度财政收入数据的季节性波动比较明显。这也表明时间序列分析中选取的时间刻度越小，越能够表现变量的变动情况。

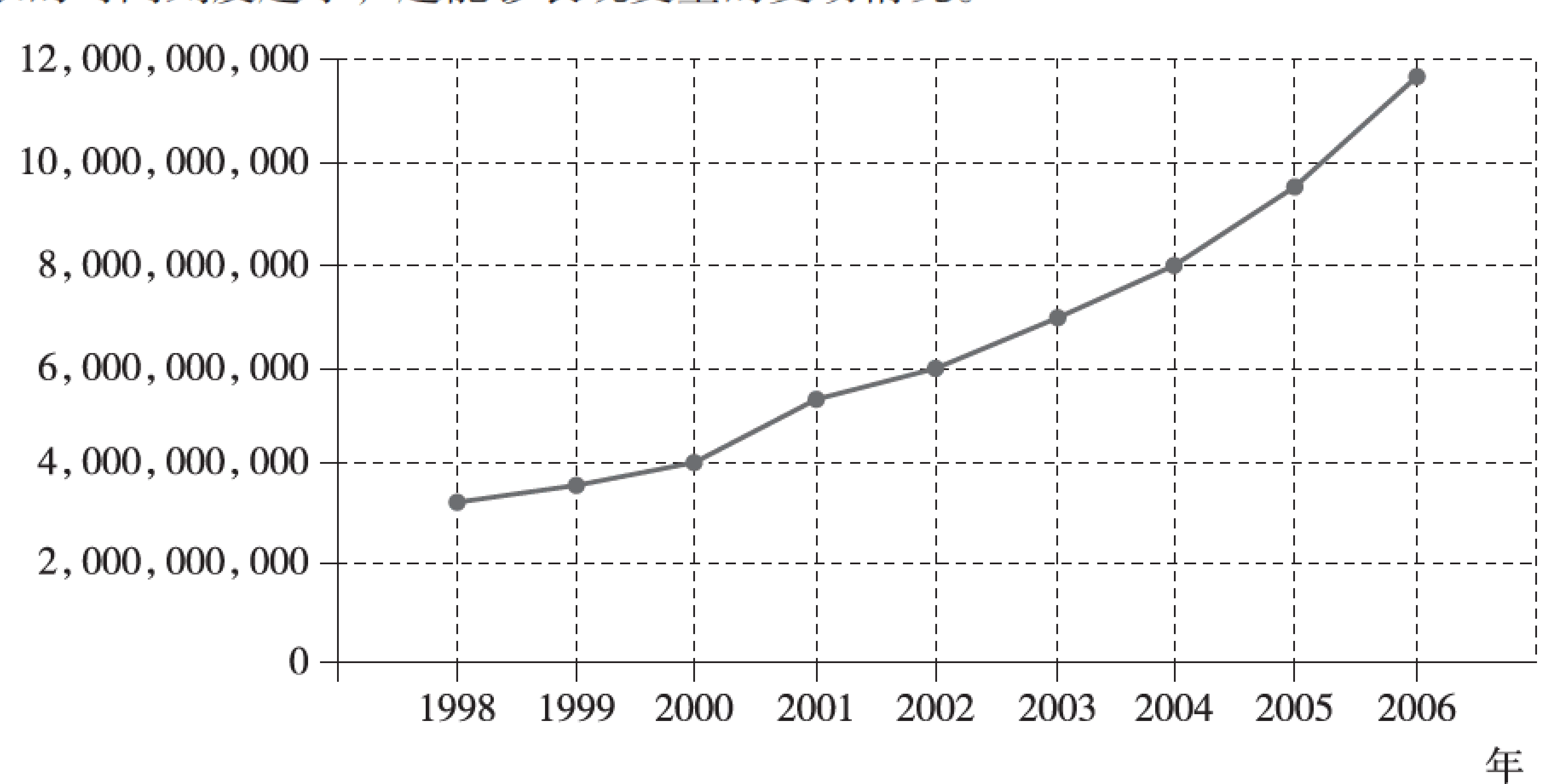


图5.10 年度时间序列数据图

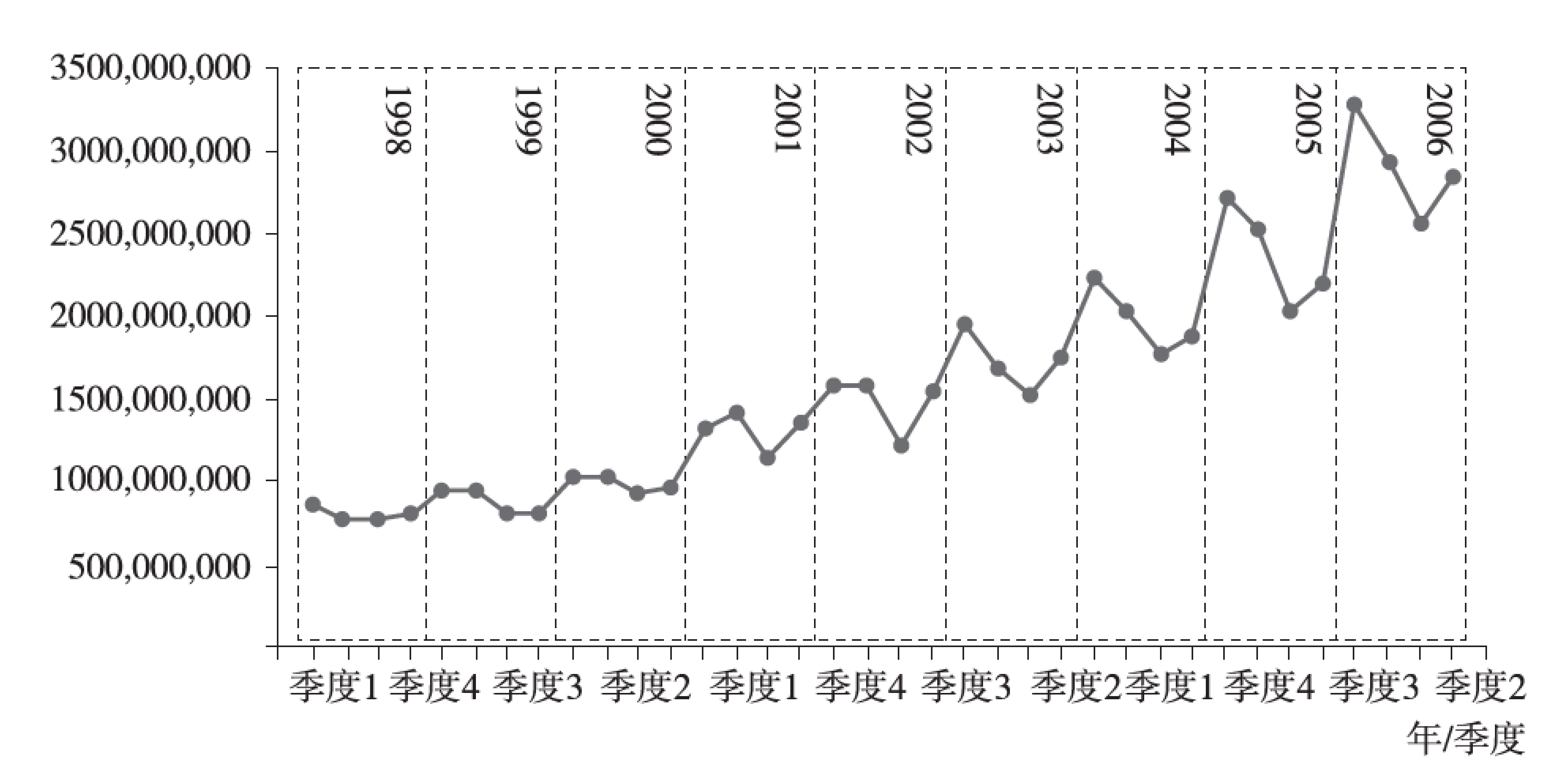


图5.11　季度时间序列数据图

统计时序分析包括两种分析方法：频域分析方法和时域分析方法。频域分析方法的原理是假设任何一种无趋势的时间序列都可以分解为若干个不同频率的周期波动，它重点分析的是频率特征，常用于电力以及工程等方面。时域分析方法的原理是事件的发展经常都具有一定的惯性，用统计的语言来描述，这种惯性就是序列值之间存在一定的相关关系，在这种相关关系中一般就含有某种统计规律。这种方法重点分析的是事物随时间的发展变迁的趋势，常用于人口、经济、气象等方面。

时序分析方法的目的是找出时间序列值之间相关关系的统计规律，并建立适当的数学模型来描述这种规律，进而利用这个模型来预测未来的走势。对于时间序列数据趋势变化可以用（按年/月）柱型图、折线图来展现，也可以利用指数平滑等技术对折线图进行趋势拟合。

（2）时间序列分析建模与应用

时间序列分析模型的建立主要有两种：曲线拟合和参数估计。目前主要建立两种类型的模型，即ARMA模型和ARIMA模型①。

ARMA模型的全称是自回归移动平均（Auto Regression Moving Average）模型，它是目前最常用的拟合平稳序列的模型，可细分为AR模型（Auto Regression Model）、MA模型（Moving Average Model）和混合ARMA模型三大类。

ARIMA模型又称自回归求和移动平均模型，当时间序列本身不是平稳的时候，如果它的增量，即一次差分，稳定在零点附近，则可以将其看成是平稳序列。在实际的应用中，所遇到的多数非平稳序列可以通过一次或多次差分后成为平稳时间序列，进而可以建立模型。这说明任何非平稳序列只要通过适当阶数的差分运算实现差分后平稳，就可以对差分后序列进行ARIMA模型拟合了。

建立时间序列模型的一般步骤如下。

①收集数据。时间序列分析法分析的是被观测系统的时间序列动态数据，对这些数据的收集，常用的方法包括观测、调查、统计和抽样等。

②作图分析。针对需要分析的时间序列动态数据，运用相关的方法制作相关图，并进行相关分析，从而得到自相关函数。相关图可以提供很多关于这些动态数据的信息，比如在图

① http://www.cnblogs.com/emanlee/archive/2012/02/06/2339650.html.

上可以观察到变化的趋势和周期，跳点以及拐点。所谓跳点，指的是那些和其他数据不一致的对象；而拐点指的是时间序列趋势骤然改变的点。对跳点和拐点的处理是不一样的。对于跳点，首先要判断是否属于正常现象，若正常，则在建立模型的过程中就要考虑到；若判断为异常现象，则需要对其进行调整，使其达到期望值。对于拐点，若其存在，则在建模的过程中，需要对时间序列进行分段拟合。分段拟合的模型是不相同的，比如采用门限回归模型。

③曲线拟合。根据分析结果，选择适当的随机模型对时间序列数据进行拟合，也就是曲线拟合。随机模型需要针对时间序列的具体情况来选择，比如，较短的或者相对比较简单的时间序列，可以选择趋势模型和季节模型并加上误差来拟合；而对于平稳时间序列，可以选择通用ARIMA模型及其特殊情况的自回归模型、滑动平均模型或组合ARIMA模型等模型。对于观测值数目较大（多于50个），可以选择ARIMA模型。对于非平稳时间序列，首先需要对其进行处理，即通过差分运算转化为平稳时间序列，得到的是差分序列，然后选择合适的模型来拟合。

时间序列是一种特殊的随机过程，当$X(t)$中的t值取非负整数时，就可以代表各个时刻，可以看作是时间序列。因此，当一个随机过程可以看作时间序列时，就可以利用现有的时间序列模型建模分析该随机过程的特性。

时间序列分析有很多实际的应用，比如气象、水文、地震以及其他自然灾害的预报，国民经济宏观控制，环境污染控制，区域综合发展规划，企业经营管理，市场潜量预测等。该方法在生态平衡、天文学和海洋学等方面也有着广泛的应用。

2. 相关分析

相关分析是一种研究变量间相关性的统计方法，包括变量间是否有依存关系，如果有是怎样的关系，以及这种依存关系的相关方向和相关程度等。它是进行因果分析的基本工具，通过相关分析可以判断经济指标之间的替代关系和关联度①。

相关分析用来研究两个变量（x，y）的相互关系，测定它们联系的紧密程度。测定的方法可以从散点图直观地进行观察，也可以通过计算相关系数r得到较为精确的判断。计算公式：

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-(\sum x)^2\times\sqrt{n\sum y^2-(\sum y)^2}}} \tag{5-3}$$

注：相关分析也可以通过某些软件，如SPSS、Eviews等计算得到。

相关分析的分类如下。

- 线性相关分析：主要针对两个变量之间存在的线性关系进行研究分析，变量间线性关系强弱的程度通常用相关系数r来刻画。如果x，y变化的方向一致，且得到$r>0$；则表示这两个变量呈正相关；若$r<0$，则表示这两个变量呈负相关；若$r=0$，则表示这两个变量无线性相关。相关系数r的取值范围是$|r|<1$，如果$|r|>0.80$时具有强的正（负）相关关系，如果$0.3<|r|<0.80$时具有弱的正（负）相关关系。如果$|r|<0.30$时

① http://baike.baidu.com/view/325793.htm?fr=aladdin.

认为没有有效的相关关系。

- 偏相关分析：该分析就是先控制一些对两变量间的相关性可能有影响的其他变量，再对两变量间的线性相关性进行研究分析。例如通过控制年龄和工作经验的影响，然后估计工资收入与受教育水平之间的相关关系。
- 距离分析：是通过距离的大小对观测量之间或变量之间相似或不相似程度的一种测度，是一种广义的距离。分为观测量之间距离分析和变量之间距离分析。

值得注意的是，相关分析并不是因果分析，不会对两个变量的因果关系进行判断，在回归分析中更强调的是自变量和随之而变的因变量。相关系数的计算方法是以直线关系为前提的，如果是曲线关系，则相关系数方法计算时会出现错误的结果。相关分析和回归分析这两种方法虽然分析的是变量之间的关系，但是相关分析是回归分析的基础，而回归分析则是认识变量之间相关程度的具体形式。

3. 回归分析

回归分析是在掌握大量观察数据的基础上，利用数理统计方法建立因变量与自变量之间的回归关系函数表达式，即回归方程式。该方法是一种统计学上分析数据的方法，主要用于得到变量之间的关系，比如这些变量是否相关联、相关的方向以及相关的强度等，同时建立相应的数学模型，从而把握特定的变量，并且对研究人员感兴趣的变量进行预测。找出一条最能够代表所有观测资料的函数曲线（回归估计式）是回归分析的目的所在。然后用此函数表示因变量和自变量之间的关系。

进行回归分析的步骤如下。

①变量的确定。变量的确定包括自变量和因变量，关键需要根据预测目标进行变量的确定。例如，如果要对下季度的商品销售量进行预测，即确定了预测的目标为销售量，则这个预测目标就是模型的因变量。确定了因变量后，就需要选择自变量了。自变量通常是与预测目标具有一定影响的因素，一般选择主要的影响因素作为自变量。

②回归模型的建立。依据自变量和因变量的历史统计资料进行计算，在此基础上建立回归分析方程，即回归分析预测模型。首先需要运用相关的方法对这些历史资料进行处理，然后再利用最小二乘法或者极大似然法的方法建立科学合理的回归分析预测模型。

③相关分析。因为回归分析进行的是自变量和因变量之间的因果关系的分析，因此建立相应的模型进行回归分析就要求自变量和因变量之间是有某种因果关系的，否则回归分析是毫无价值的。所以，进行回归分析需要掌握自变量和因变量的一些相关信息：比如自变量和因变量之间是否存在一定的相关性，如果存在那么相关的程度是怎么样的，能否较为准确地判断相关程度等。相关系数是相关分析一个比较重要的概念，它能够作为自变量和因变量之间相关程度的度量，因此，进行相关分析通常需要确定出相关系数的具体数值。

④模型的检验，预测误差的计算。建立的回归预测模型不一定可以对实际问题进行预测，需要对模型进行检验。此外，预测误差也是衡量一个回归预测模型的重要指标。一个好的回归预测模型必须能够通过多方面的检验，同时它的预测误差也是比较小的。这样的模型可以得到比较好的预测结果。

⑤预测值的确定。运用之前确定的回归预测模型进行预测值的计算，再根据具体的实际

情况，运用相关知识进行全面分析，从而得到最终的预测值。

回归分析有很多种类，按照不同的标准可以分成不同的类别。如果回归分析模型中只有一个自变量，则为一元回归分析；如果自变量的个数多于一个，则为多元回归分析。这种划分是按照自变量的个数进行的。也可以依据自变量的次方进行划分，得到线性回归模型和非线性回归模型。线性回归分析模型中的自变量都是一次方的；非线性回归分析是除了线性回归分析以外的回归分析。下面以线性回归为例，来说明回归方程以及相关系数、回归系数的求法。

线性回归方程可以由许多相关的方法得到，通常运用最小二乘法，如$y=bx+a$的直线，可以计算出它的经验拟合方程①，计算公式：

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{5-4}$$

通过以下公式来得到相关系数，从而判断拟合的好坏程度：

$$a=y-b\bar{x} \tag{5-5}$$

对大部分的行为研究者来讲，在回归结果输出中，最重要的是回归系数，一般要求这个值大于0.05，则变量间的影响不是显著的，若小于0.05，则是显著的。

利用最小二乘法能够计算出线性回归模型中的参数，但是这样的线性回归方程若需要盈余与实际问题的分析时，还需要对回归方程的线性关系进行各种统计检验，如：回归方程显著性检验、回归系数显著性检验、残差分析等。在实际运用中，线性回归的应用非常广泛，可分为以下两大类：

- 预测功能。若分析的目标是预测或者是映射，线性回归模型能够根据观测到的数据集合以及x的值，拟合得到一个预测模型。这个模型可以针对新出现的x，在没有给定与其相应的y值的情况下，给出一个预测值y。
- 确定变量间的相关性以及相关程度。若给定一个变量y和一系列变量x_1，…，x_p，这一系列变量可能与y有关联，也可能不相关。通过线性回归分析能够得到y与这一系列的x的相关程度的一个量化，同时能够筛选出哪些变量与y相关，哪些变量与y相关，并对含有与y相关冗余信息的变量x_j的子集进行识别。

4. 判别分析

当得到一个新的样本数据时，如果要确定该样品属于已知类型中的哪一类，就需要运用判别分析来解决该问题。判别分析也称“分辨法”，是类别明确的一种分类技术，其分类方式是事先确定，依据若干变量值判断研究对象归属问题的一种多变量统计分析方法。

判别分析的基本原理是按照一定的判别原则，建立一个或多个判别函数，用研究对象的大量资料确定判别函数中的变量系数，并计算判别指标。根据判别指标就可以确定观察对象属于哪一类。其主要目的是根据已知类别的样本建立判别模型，然后依据该模型判别未知类别的样本的归属问题。

判别函数的一般公式：

$$y=a_1x_1+a_2x_2+\ldots,a_nx_n \tag{5-6}$$

① http://baike.baidu.com/view/449540.htm?fr=aladdin.

其中，Y表示判别值，x_1，x_2，…，xn为反映研究对象的变量值；a_1，a_2，…，a_n表示各变量的系数，也称判别系数。

判别分析的分类如图5.12所示。

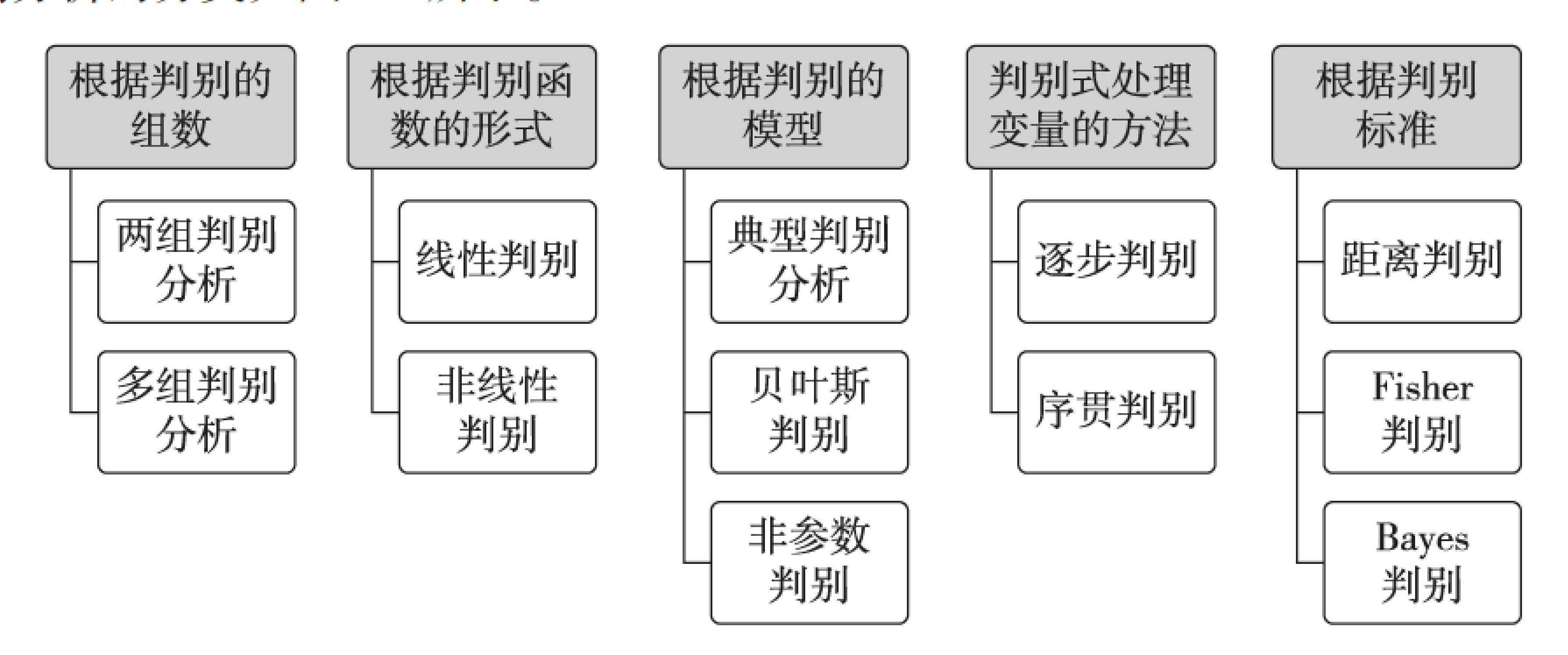


图5.12　判别分析分类图

投影是费歇（Fisher）判别的基本思想，通过投影将多维问题转化成为较简单的一维问题来分析。投影的基本步骤是首先选择投影轴，确定正确的投影轴后，再将所有的样本点投影到该投影轴上，从而得到一个投影值。投影轴的选择是有一定规定的，方向要求：保证每一个类中的投影值得到的类内离差尽量小，不同的类之间的投影值得到的类间离差尽量大。概率是贝叶斯（Bayes）判别的基本思想，即首先由先验概率计算出后验概率，然后根据后验概率的分布进行判别分析。贝叶斯（Bayes）判别涉及两个概念：先验概率和后验概率。先验概率，即对研究对象事先认识的程度用概率来描述，这个概率称为先验概率；后验概率，即由一些具体的资料、先验概率以及给定的判别规则，通过计算得到的概率。后验概率是对先验概率的修正。

距离判别的基本依据是距离，即各个样本与各个母体间的距离。距离判别就是依据这个距离进行判别的。首先需要得到各个母体的距离判别函数，这个函数式是依据具体的资料构建的。然后，依次代入各个样本数据，从而计算出各个样本与各个母体间的距离。最后，根据距离值最小原则对样本归属进行判别，即样本与哪个母体的距离最小那么就属于那个母体[①]。

5. 主成分分析与因子分析

（1）主成分分析

主成分分析（Principal Component Analysis，PCA），是一种多元统计分析方法。该方法把多个变量进行线性变换，从而得到不相关的综合变量，再根据给定的规则从中选择出少数几个能够较好反映出原始变量信息的综合变量。K.皮尔森对非随机变量最先引入了主成分分析法，后来H.霍特林将该方法推广到了随机向量，用于分析数据及建立数理模型。该方法主要是通过对协方差矩阵进行特征分解，以得出数据的主成分（即特征向量）及它们的权值（即特征值）。

主成分分析的实际常用计算步骤如下。

①计算相关系数矩阵。

②求出相关系数矩阵的特征值i及相应的正交化单位特征向量a_i。

① http://blog.sina.com.cn/s/blog_436dc2b801011qt9.html

③选择主成分。

④计算主成分得分。

（2）因子分析

因子分析是英国心理学家C.E.斯皮尔曼提出的，是一种将多变量化简的技术。该方法依据变量之间的相关性，把多个具有重叠信息、关系复杂的变量归为少数几个不相关的变量，即综合因子①。因子分析的由来是斯皮尔曼发现如果一个学生有一科成绩比较好，那么通常这位学生其余科目的成绩也会比较好，即学生的各科成绩之间存在着一定的相关性。根据这一事实，推断是否有某些共性因子，或者存在某些智力条件因素对学生的课程成绩产生影响②。

由原始变量得到少数几个代表因子是因子分析的目的，但前提要求是，变量之间应该具有较强的相关关系。它的基本思想是：按照相关性的强度对原有变量进行分组，使得不同组内的变量之间没有相关性或者相关性较低，同组的变量相关性高，从而得到共性因子也就是每组变量代表的一个基本结构。因子分析不但可以减少变量的数目，还可以检验变量间关系的假设。它主要是用于寻找变量之间的潜在结构、内在结构的证实以及评估问卷的结构效度等。根据方法，因子分析可以分为探索性因子分析和验证性因子分析，它们之间的主要区别是有没有事先假定因子与测度项之间的关系。

（3）因子分析的数学模型

因子分析的过程是将许多原有的变量表示成几个具有代表性的公共因子。下面，利用数学模型来表示其主要过程。因子分析模型为：

$$
\begin{aligned}
x_1 &= a_{11}F_1 + a_{12}F_2 + \cdots a_{1m}F_m + \varepsilon_1 \\
x_2 &= a_{21}F_1 + a_{22}F_2 + \cdots a_{1p}F_p + \varepsilon_2 \\
&\vdots \\
x_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots a_{pm}F_m + \varepsilon_p
\end{aligned}
\tag{5-7}
$$

说明：a_{ij}，因子载荷（实际上是权数）；F_i，公共因子（主因子）；$a_{11} \sim a_{mp}$，组成因子载荷矩阵；ε_i，特殊因子。

因子载荷的统计意义是，第i个变量与第j个公共因子的相关系数，即表示变量X_i依赖于F_j的比重，心理学家将它称为载荷。特殊因子表示原始变量中不能由因子解释的部分，均值为0。

（4）主成分分析与因子分析的区别

主成分分析和因子分析无论从基本思想上还是应用上都有相似之处，但也存在一些区别，可以总结为以下七点。

- 原理不同。主成分分析基本原理是利用降维（线性变换）的思想，从多个原始变量提取几个不相关的主成分，使得提取出来的主成分与原始变量相比具有某些更优越的性能，从而达到既不忽视问题的实质，又简化了系统结构的目的。因子分析基本原理也是利用降维的思想，提取几个支配原始变量的公因子和一个特殊因子，各个公因子之间可以是相关或不相关的。
- 线性表示方向不同。因子分析是用各个公因子的线性组合来表示变量，而主成分分析

① http://blog.sina.com.cn/s/blog_6b36e67501013mmd.html

② http://www.docin.com/p-30447979.html.

则是用主成分的线性组合表示各变量。

- 假设条件不同。进行因子分析要做一定的假设，这些假设为各共性因子不相关、各特殊因子（specificfactor）不相关、共性因子与特殊因子不相关等；而进行主成分分析不用做任何假设[①]。
- 求解方法不同。求解主成分的方法有从协方差阵出发（协方差阵已知）和从相关阵出发（相关阵R已知），采用的方法只有主成分分法。求解因子载荷的方法有很多种，包括主成分分法、主轴因子法、极大似然法、最小二乘法、a因子提取法等。
- 解释重点不同。因子分析提取的共性因子比主成分分析提取的主成分更具有解释性。主成分分析把解释的重点放在各变量的总方差，忽略各变量的度量误差；而因子分析把解释的重点放在各变量间的协方差，潜在变量修正了观察变量的误差。
- 因子和主成分的变化不同。如果相关矩阵的特征值或协方差矩阵惟一，那么主成分通常是固定的；而由于因子可以旋转得到不同的因子，所以因子不是固定不变的。
- 数量的不同。主成分的个数是固定的，通常情况下主成分的数目和变量数是相同的，不同的只是主成分解释的信息量；而因子的数目是用户给定的，比如运用SPSS软件时，可以根据相应的条件自动设定，只要是特征值大于1的因子便进入分析。指定的因子数目影响分析的结果，因子数目不同，结果也不同。

6. 对应分析

对应分析（Correspondence Analysis）也称关联分析、R-Q型因子分析，是近年新发展起来的一种多元统计分析技术，对由定性变量构成的交互汇总表进行分析，以此来揭示变量间的联系。主要适用于有多个类别的定类变量，可以揭示同一变量的各个类别之间的差异，以及不同变量各个类别之间的对应关系。对应分析的基本思想是将一个列联表的行和列中各元素的比例结构以点的形式在较低维的空间中表示出来。其最大的特点是能把众多的样本和变量同时展现在同一张图上，将样本的大类及其属性在图上直观而又明了地表示出来，具有直观性。因此，它是强有力的数据图示化技术，当然也是强有力的市场研究分析技术。

对应分析法整个处理过程由两部分组成：表格和关联图。表格在对应分析法中指的是一个二维的表格，它由行和列组成。每行代表研究对象的一个属性，依次排开。列则代表不同的对象本身，它由样本集合构成，排列顺序并没有特别的要求。在关联图上，各个样本都浓缩为一个点集合，而样本的属性变量在图上同样也是以点集合的形式展现出来。

对应分析对数据的格式要求：

- 对应分析数据的典型格式是列联表和交叉频数表。
- 常表示不同背景的消费者对若干产品或产品属性的选择频率。
- 背景变量或属性变量可以单独使用或并列使用。
- 两个变量间称为简单对应分析。
- 多个变量间称为多元对应分析。

例如，假定有个汽车数据集，包括：来源国（1—美国、2—欧洲、3—日本），尺寸

① http://wenku.baidu.com/view/18d600d284254b35eefd34de.html

（1—大型、2—中型、3—小型），类型（1—家庭、2—运动、3—工作），拥有（1—自有、2—租赁），性别（1—男、2—女），收入来源（1—1份工资来源、2—2份工资来源），婚姻状况（1—已婚、2—已婚有孩子、3—单身、4—单身有孩子）。

从数据集看，有7个定类变量，当组合成简单的交叉表是困难的事情时，就应采用多重对应分析。本例采用SPSS来对其进行对应分析，可以把要分析的变量都放到分析变量内，选择类别图（每一个变量的分类图，重点是选择联合类别图）。当把7个变量全部放入，然后执行，之后得到如图5.13所示的结果。

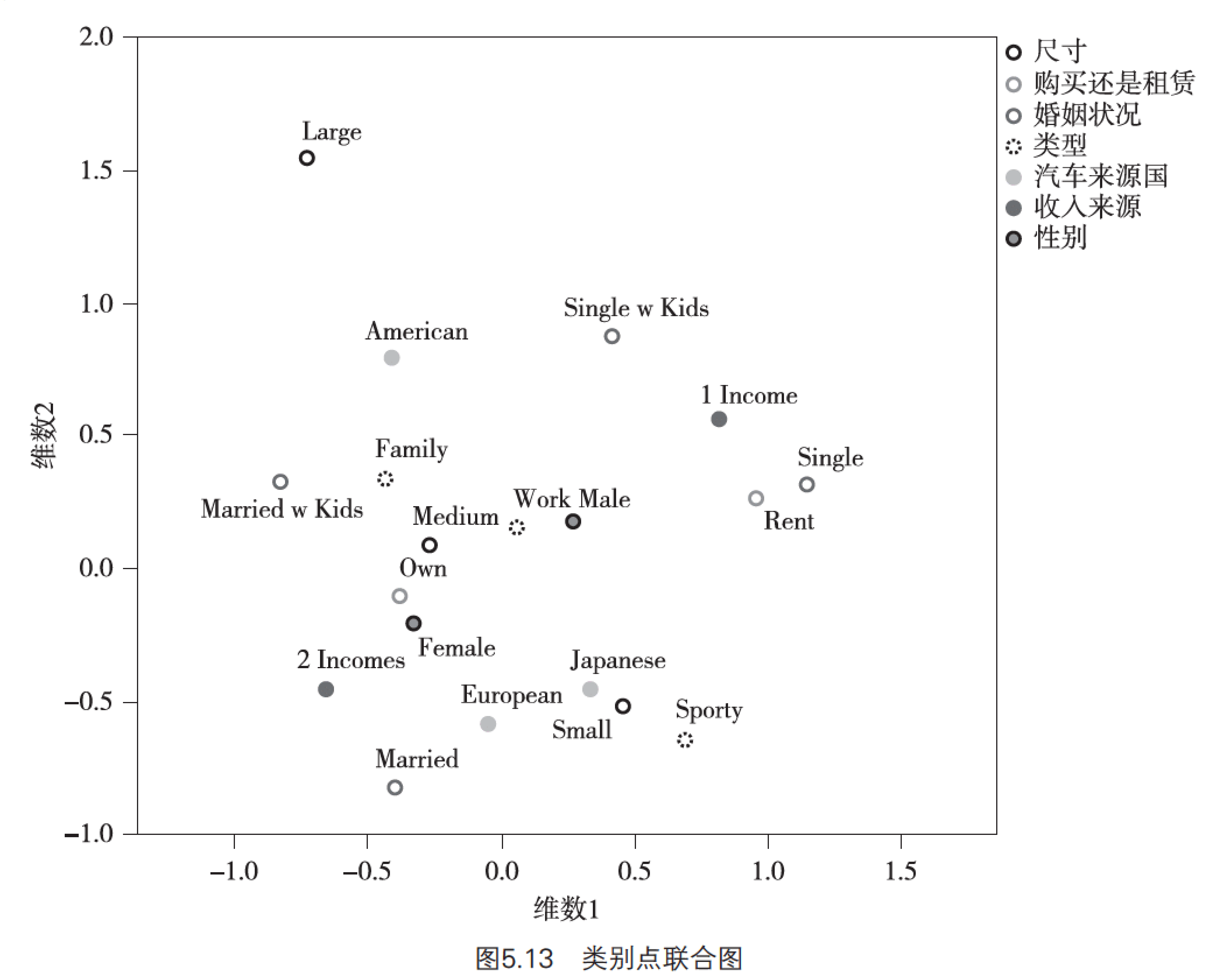


图5.13　类别点联合图

从图5.13中可以看出：美国车都比较大，家庭型，主要购买者是已婚带孩子的；日本和欧洲车主要是小型、运动的，主要购买者是已婚没有孩子的人购买；值得注意的是单身和单身带孩子的往往选择租赁汽车，因为这类人收入来源单一。但是这个地区没有车满足这个市场，也许是市场空白。

对应分析主要应用领域：概念发展（Concept Development）、新产品开发（New Product Development）、市场细分（Market Segmentation）、竞争分析（Competitive Analysis）和广告研究（Advertisement Research）等。

7. 多维尺度分析

多维尺度分析（Multi-dimension Scaling，MDS）一种将多维空间的研究对象（样本或变量）简化到低维空间（一般是二维到三维空间）进行定位、分析和归类，同时又保留对象间

原始关系的数据分析方法。比如在消费者行为分析上，可以将消费者对品牌的感觉偏好，以点的形式反映在多维空间上，而对不同品牌的感觉或偏好的差异程度，则是通过点与点间的距离体现的。这种品牌或项目的空间定位点图称为空间图，空间轴代表着消费者得以形成对品牌的感觉或偏好的各种因素或变量。多维尺度分析是一种探索性数据分析方法。该方法的基本思想是被访者对研究对象相似性的感知，是用被访者对研究对象的分组来反映的。多维尺度分析是分析消费者感觉和偏好的最有效的方法，它能以直观图的方式提供一个简化的分析方法，具有一定直观性和合理性。

距离矩阵通常可以使用两种方法得到，直接相似性评价法和间接评价法。直接相似性评价法的操作步骤为：首先把评价对象两两组合，再组合被访问者，从而直接进行相似性评价。间接评价法的步骤是：研究人员先根据之前的经验发现对评价影响比较大的属性，再由被访问者评价这些研究对象的属性，最后用多维空间的坐标来表示所有的属性，从而计算出对象间的距离。该距离通常运用距离变换进行计算[①]。

多维尺度分析有很多方面的应用，比如品牌形象评价。品牌形象评价的基本思路是把握消费者对本公司和竞争者的品牌认知的差异程度，根据差异的程度了解相对于竞争对手本公司在消费者心中的位置。例如，针对广州市民对市内各医院的评价，包括对专业、服务、费用、方便等方面的感知评价，利用多维尺度分析得到空间定位图，如图5.14所示[②]。

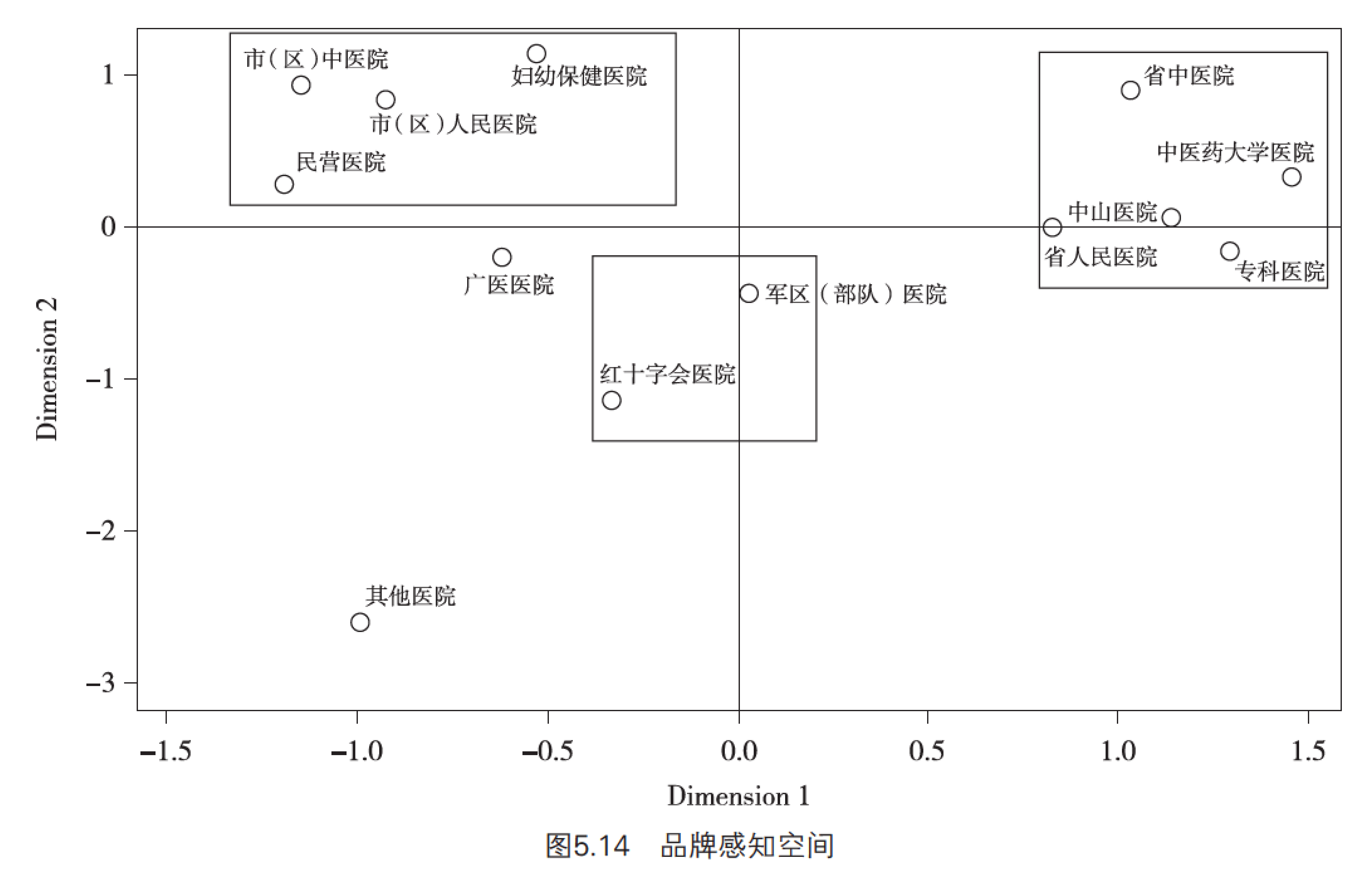


图5.14　品牌感知空间

结果表明，广州市民对市内各医院的感知评价大致可以分为三类：第一类属于专业性强、技术高的医院，包括中山医院、省人民医院、省中医院、中医药大学医院及专科医院；第二类属于费用比较合理的医院，包括市（区）的中医院、人民医院及妇幼保健医院；第三

① http://wenku.baidu.com/view/18d600d284254b35eefd34de.html

② http://blog.sina.com.cn/s/blog_6ce00d7b0100xasj.html

类则为特点不明显类，包括红十字会医院、军区（部队）医院（注：由于样本数量限制，分院、同类型医院合并分析，差异性有所平均，结论仅供参考）。

运用对应分析来分析民众对医院的感知评价，需要对数据进行处理。首先，把以行和列为变量的交叉表转换为对应分析图，再用各散点空间位置关系的形式表现表格中包含的类别关联信息。上述数据运用对应分析后呈现如图5.15所示[①]的结果。

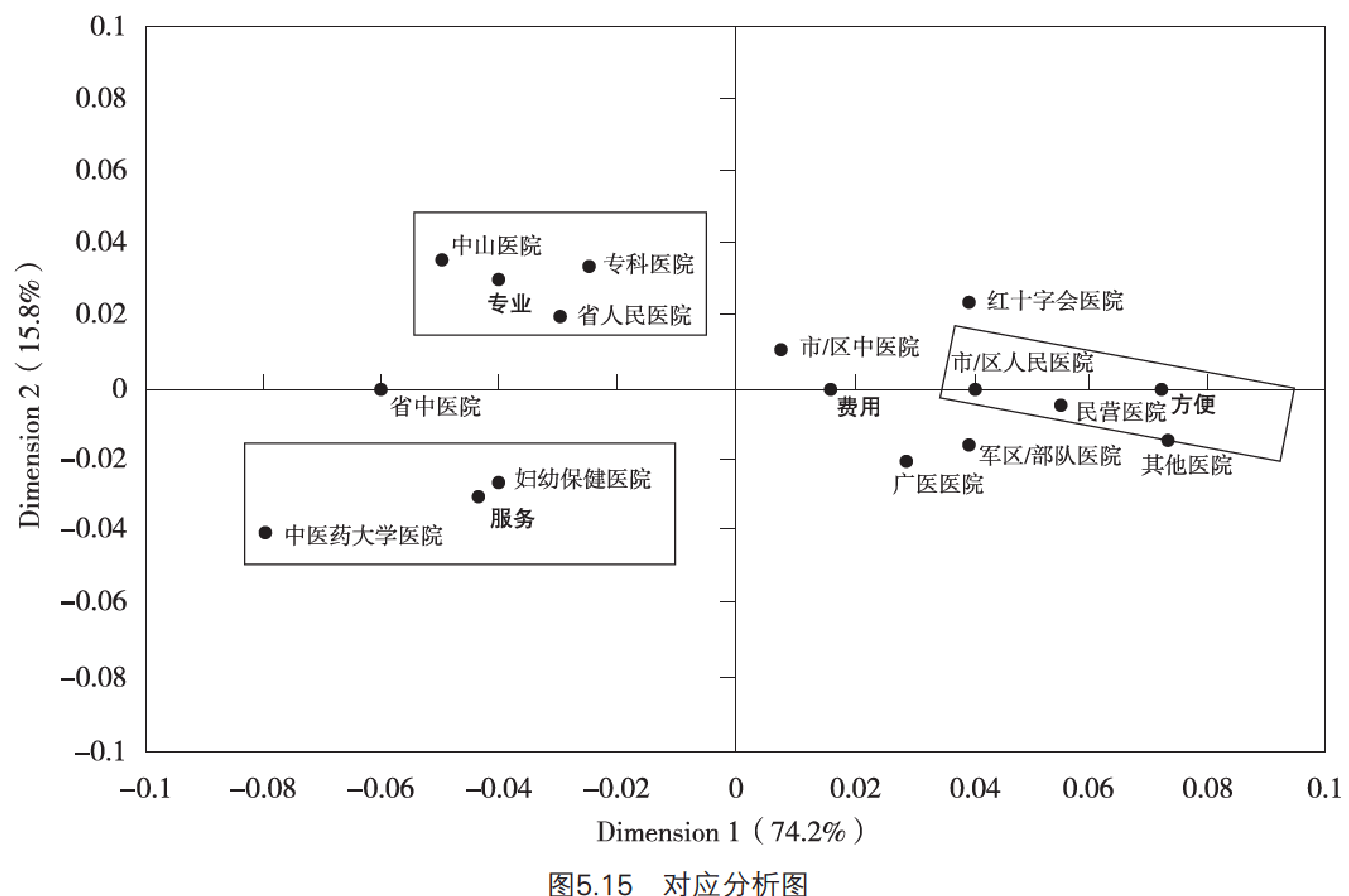


图5.15　对应分析图

观察上述两种方法的结果分析图，可以发现它们的区别：多维尺度空间图展现的是各医院间的相对位置；而对应分析图展现的是各医院与医院专业、服务、费用、方便等方面的相关性。另外，通过这两种方法的对比发现，多维尺度分析得到的是行变量间的相似性或者差异性，也就是各所医院间的相似性或差异性。而对应分析得到的是行变量与列变量之间的相关性，比如行变量“中山医院”和列变量“医院专业水平、医院服务、医院费用等”之间的相关性。也就是说，多维尺度分析强调的是行变量间的关系，对应分析强调的是行变量与列变量之间的关系。

5.2.3　可视化技术

在1987年根据美国国家科学基金会召开的“科学计算可视化研讨会”内容撰写的一份报告中正式提出了“可视化”一词，其全称是“科学计算可视化”（Visualization in Scientific Computing，缩写为ViSC）。数据可视化是指对大型数据仓库或者数据库的数据进行探索，从而用较为直观的、以图像或图形的形式来表示数据所产生的信息。这样，使得用户不仅能通过关系型数据库来观察和分析数据，更可以通过直观的方式看到数据及其结构关系。数据可

① http://blog.sina.com.cn/s/blog_6ce00d7b0100xasj.html

视化的基本思想是将数据库中的数据项作为图元元素，然后把大量的数据集构成数据图形，以使用户对数据的观察和分析更深入。

数据可视化的处理对象是数据，它包含两个分支：处理科学数据的科学可视化与处理抽象的、非结构化信息的信息可视化。作为数据内涵信息的展示方法和人机交互接口，数据可视化已成为数据科学的核心要素之一，面对大规模数据，很多时候不可能通过直接观察数据本身或对数据进行简单地统计分析后就能得到数据中蕴涵的信息。例如，无法直接通过查看海量的服务器日志来判断海量数据，但是可以通过可视化将其变成形象生动的图形，这有助于对数据中的属性、关系进行深入地研究。可视化不仅可以作用于数据科学过程中不同的部分，也可以作为一种人机交互手段，贯穿于整个数据处理过程，如图5.16所示。

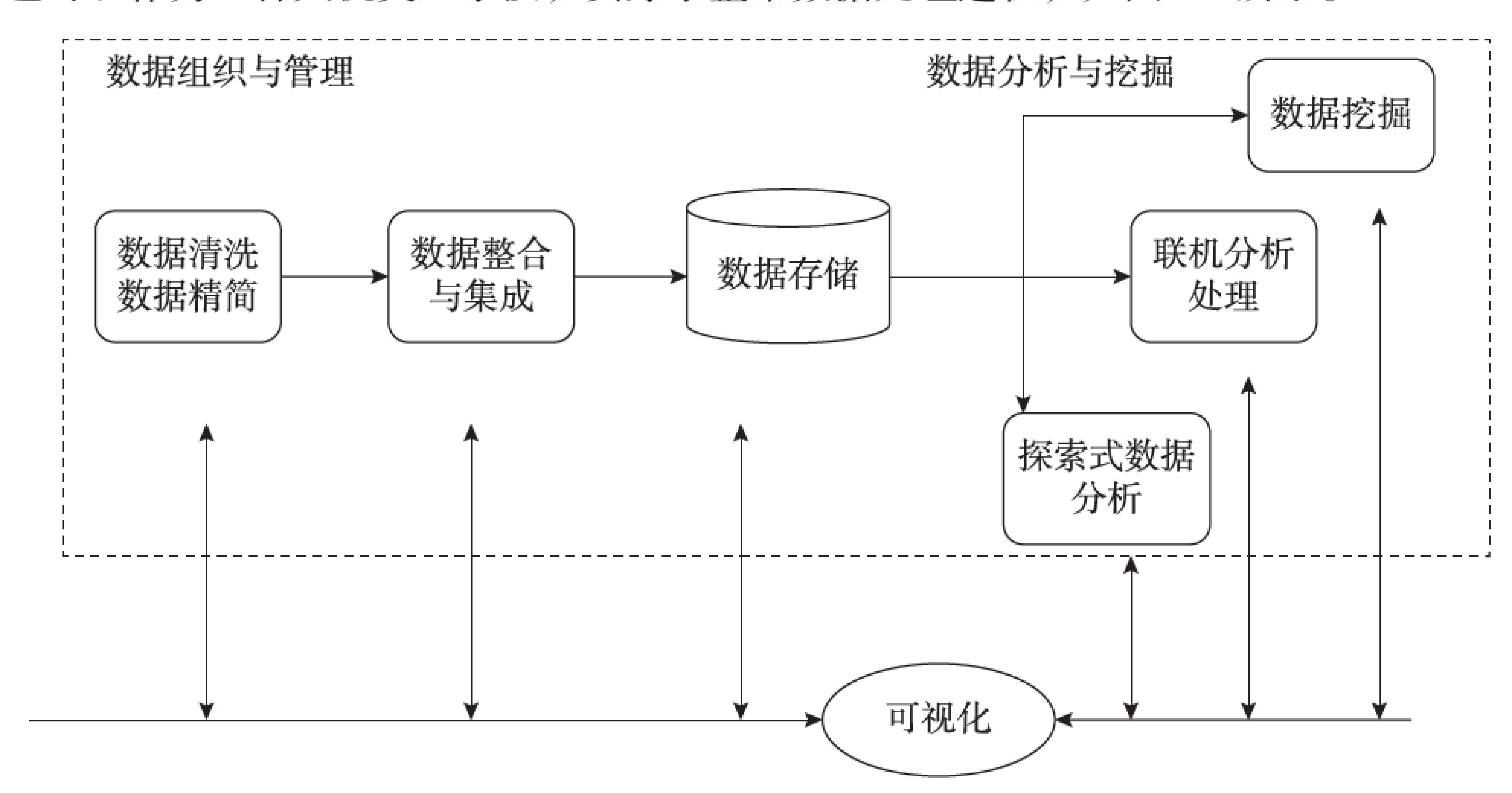


图5.16　可视化作为人机交互手段

数据可视化技术是利用计算机的巨大处理能力及计算机图像和图形学中的基本算法把海量的数据转换为静态或动态的图像或图形，呈现在人们的面前。数据可视化技术允许通过人机交互手段控制数据的提取和画面的显示，挖掘出藏于数据背后不可见的现象，为人们分析数据、理解数据、形成概念、找出规律提供强有力的手段。

数据可视化技术主要有三个特点：

- 交互性。用户可以方便地以交互的方式管理和开发数据。
- 多维性。可以看到表示对象或事件的数据的多个属性或变量，而数据可以按其每一维的值，将其分类、排序、组合和显示。
- 可视性。利用图像、曲线、二维图形、三维体和动画来显示数据，并对其模式和相互关系进行可视化分析。

数据可视化技术包含以下几个基本概念[①]：

- 数据空间。由n维属性和m个元素组成的数据集所构成的多维信息空间。
- 数据开发。指利用一定的算法和工具对数据进行定量的计算和推演。
- 数据分析。指对多维数据进行切片、分块、旋转等动作来剖析数据，从而能从多角度

① http://baike.baidu.com/view/69231.htm?fr=aladdin

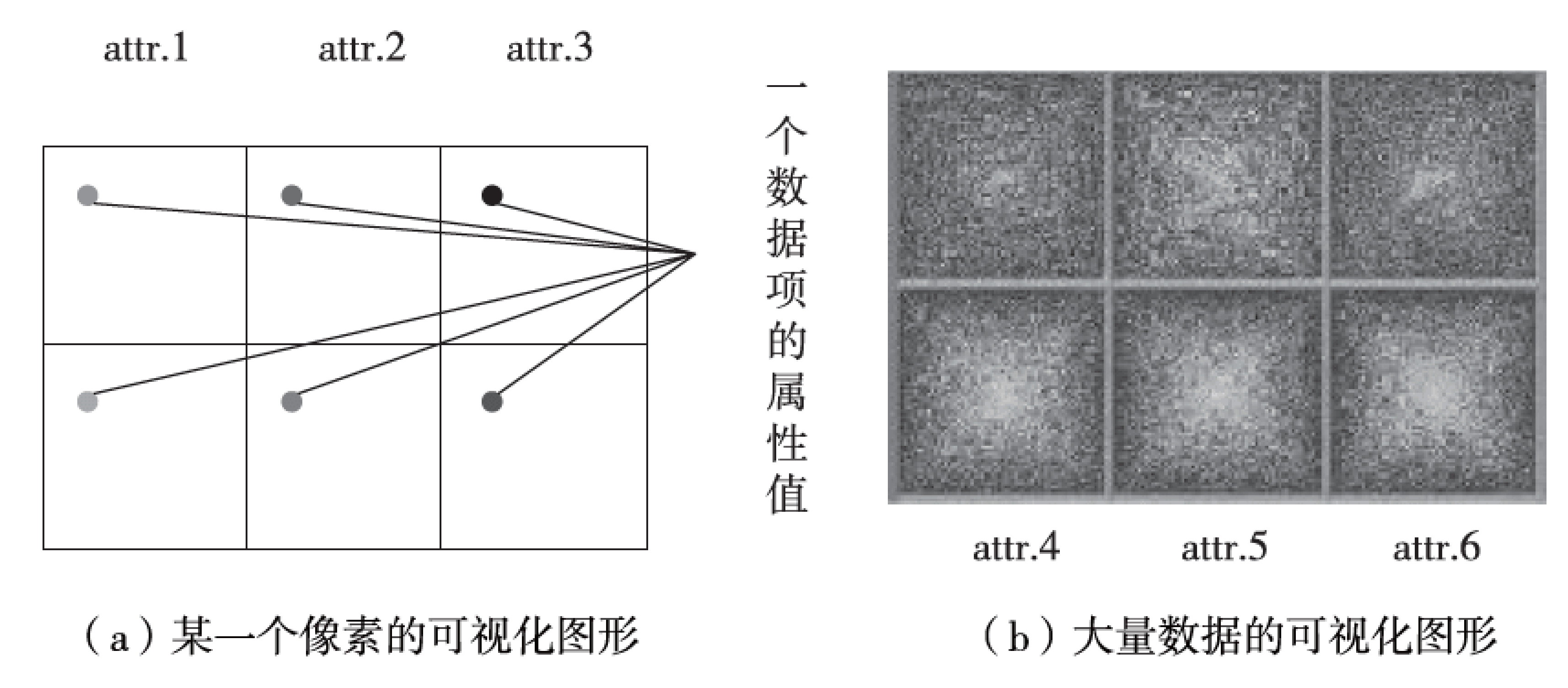
多侧面来观察数据。

- 数据可视化。指将大型数据集中的数据以图形或图像形式展现出来，并利用数据分析和开发工具来发现其中隐藏信息的处理过程。

除了很多读者已经熟悉的一些基本的显示技术，如：折线图、柱形图、散点图、条形图、面积图、圆环图以及曲面图、股价图等，随着科技的发展，针对数据可视化又出现了更多的新技术。根据其可视化的原理不同，可以把这些技术划分为基于几何的技术、基于图标的技术、面向像素的技术、基于层次的技术、基于图像的技术和分布式技术等类型。下面介绍其中几项主要的技术。

1. 面向像素技术

面向像素技术（pixel-oriented techniques）是由D.A.Keim提出的，并且开发了VisDB可视化系统。面向像素技术作为海量高维数据可视化技术，其基本思想是每一个数据项的属性映射成一个带颜色的屏幕像素，对于不同的数据属性分别以不同的窗口表示，如图5.17所示。能在屏幕中尽可能多地显示出相关的数据项是面向像素技术主要的特点，对于高分辨率的显示器来说，显示的数据最多可以达到10^6数量级。



（a）某一个像素的可视化图形　（b）大量数据的可视化图形

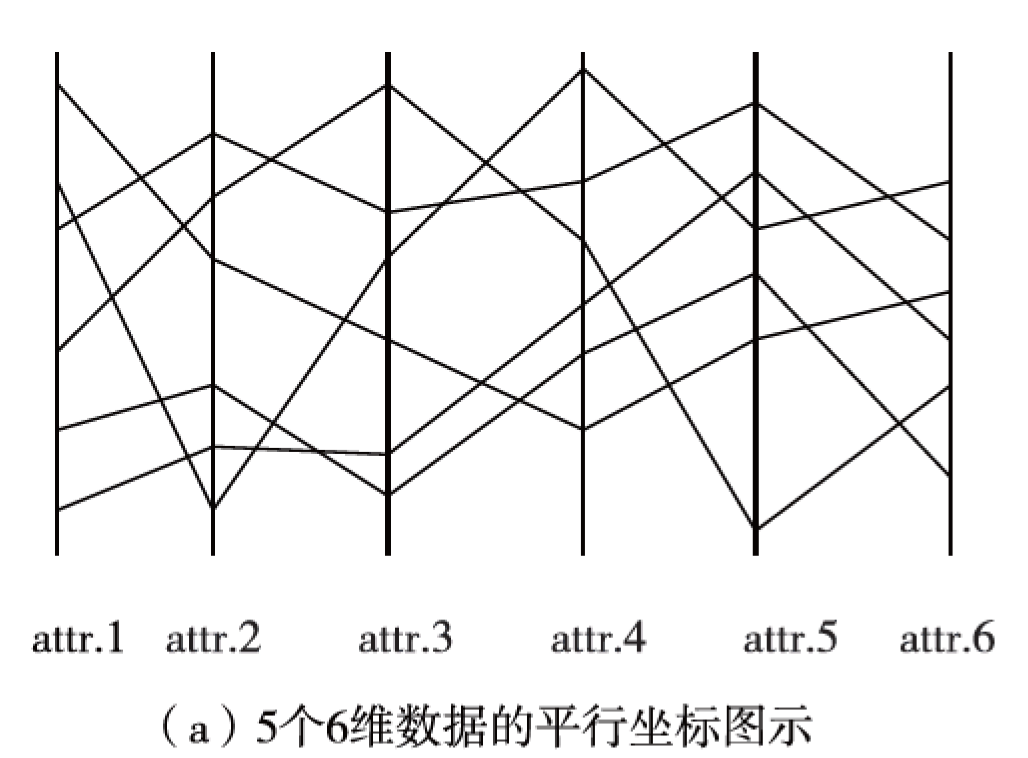
图5.17　可视化6维数据集

面向像素的可视化方法有两种：独立于查询的方法和基于查询的方法。独立于查询的方法是将数据库中的数据按照从上到下（从左到右）的次序一列一列（一行一行）地排列显示出来，数据值的变化范围与事先固定好的颜色变化范围相对应；基于查询的方法是根据数据值和所查询的要求的符合程度来匹配不同的颜色。针对每一个数据项的值（$a_1, a_2, \cdots, a_n$）及查询要求（$q_1, q_2, \cdots, q_n$）通过一个距离函数Distances（$d_1, d_2, \cdots, d_n$）计算每个属性值与查询要求的匹配值，得到每个数据的一个总的距离值d_{n+1}，以反映数据项与查询要求之间的匹配程度。总的距离值d_{n+1}越小，说明越是用户所希望看到的数据。按d_{n+1}的值由小到大把查询的数据结果从屏幕的中央螺旋地向四周展开。这样不仅能看到要查询的数据，而且可以直观地了解数据从近似匹配到不匹配的走势。但是面向像素技术存在一些缺点，如大量像素点的不规则分布，不容易进行聚类分析等。

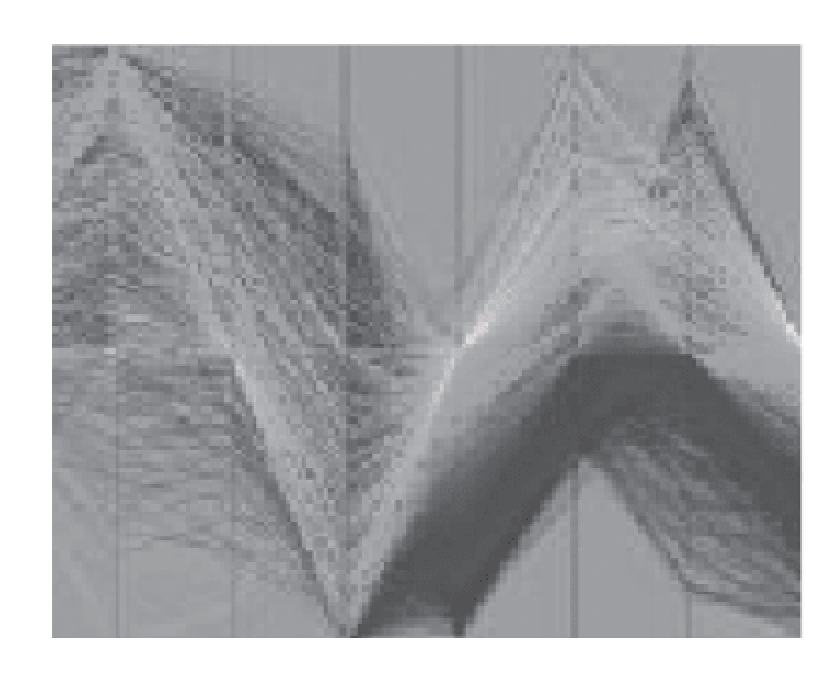
2. 基于几何的技术

目前，基于几何（geometric techniques）的可视化技术主要包括：散点矩阵、投影—截面组合视图技术、地形图、多维切片和平行坐标，它通过几何画法或几何投影技术来展现

数据库中的数据，通过线或者折线来表示数据各变量的联系。其中最早提出以二维形式表示n维数据的可视化技术之一是平行坐标。平行坐标主要用于对高维几何和多元数据的可视化，其基本思想是使用相互平行且等距的坐标轴将K维空间降维映射成两维来显示。每一根折线表示一个数据项，从图5.18可以看到折线与每个坐标轴都有一个交点，这就是该数据项在维上对应的值。



（a）5个6维数据的平行坐标图示



（b）平行坐标法开发的大型数据集

图5.18　平行坐标示例

可以借助平行坐标法开发系统，具有代表性的系统包括Parallel Visual Explorer（IBM）、XMDV（Matt Ward）、AVS/Express（van Wijk）等。其最大的优点是具有良好的数学基础，而且射影几何解释和对偶特性有助于进行可视化数据分析。对于大型的数据集它能反映出各维属性间的关系以及数据在各维属性之间的走向趋势，对于小型的数据集用户能通过二维平面看到所有数据的n维属性。平行坐标的优点是表达的维数决定于屏幕的水平宽度，而不需要使用矢量或者其他可视图标。但是大量的交迭线使得折线密度增加，层次不清，用户很难识别，此时可以通过层次的方式组织数据集，引入交互手段，以分层显示的方法来解决这一问题。

3. 基于图标的技术

基于图标（icon-based techniques）技术又称图标显示技术，它的基本思想是定制一些称为图标（矩阵、椎体、箭头）的几何对象，然后将每一个多维数据项映射成为一个对应的图标，并按照一定顺序排列这些图标。图标的各项属性，如大小、颜色、形状等均对应于数据项的维。基于图标的可视化技术包括枝形图、形状编码、彩色图标等，这种技术适用于那些在二维平面上具有良好展开属性的数据集。

枝形图方法（Stick Figures）是其中的基本方法之一，它的基本思想是用同一棵树的枝表示多个变量，每一个树枝表示一个变量。枝形图首先选取多维属性中的两种属性作为基本的X-Y平面坐标轴，在此平面上利用小树枝的长度或角度的不同代表其他属性值的变化。例如图5.19所示的两个数据点，它们的左边的二维属性含有相同的数据值，而右边的二维属性的数据值则不相同。

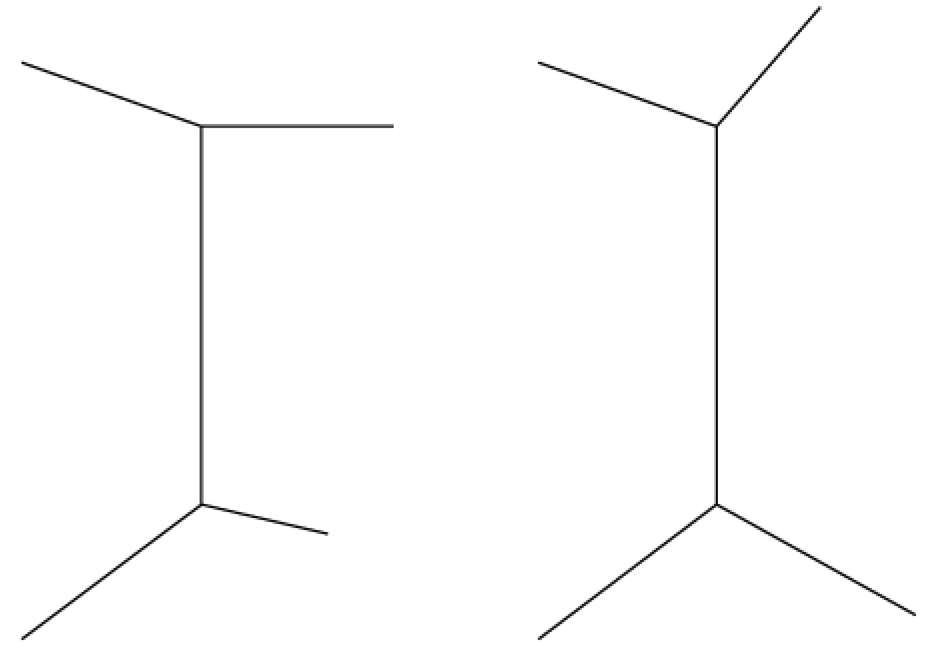


图5.19　树枝图法两个数据点的表示

4. 基于层次的可视化技术

基于层次（hierarchinal techniques）的可视化技术主要针对数据库系统中具有层次结构的数据信息，它的基本思想是将k维数据空间划分为若干子空间，对这些子空间仍以层次结构的方式组织并通过图形方式表示出来。基于层次的可视化方法多利用树形结构，这样可以直接应用于具有层次结构的数据，也可以对数据变量进行层次划分，不同层次表示不同变量值。基于层次的可视化技术包括多维堆积图、树图、锥型树等方法。树图（Treemap）是其中的一种代表技术。把图5.20（a）所示的树形结构数据在图5.20（b）中以树图的形式表示出来。图中每一个结点都有名称和数值大小，父结点是各个子结点大小的总和。

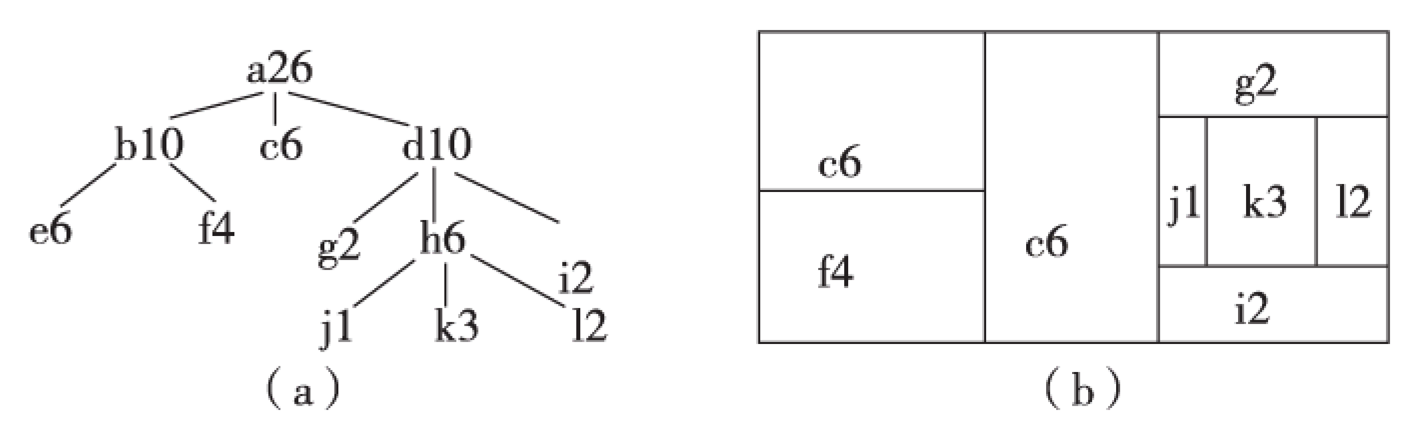


图5.20 一个层次结构的树图显示范例

树图是可视化层次结构数据的重要方法，适合于大量的层次数据集。它的基本思想是依据数据的层次结构将屏幕空间划分成若干个矩形子空间，子空间大小由结点大小决定；树图层次则依据根结点到叶结点的顺序，沿水平和垂直依次转换，开始将空间水平划分，下一层将得到的子空间垂直划分，再下一层又是水平划分，依此类推。树图允许对每一个划分的矩形进行相应的颜色匹配或给予必要的说明。

基于层次的可视化技术开发的系统主要有：Hyperbolic Trees（Xerox）、Info Cube（Sony）、Elastic Windows（HCIL-Maryland）、TreeMap（HCIL-Maryland）等。

5. 其他的数据可视化技术

数据可视化还有很多其他的技术和方法。比如，HD-Eye算法结合数据挖掘技术，先将数据分簇，再对感兴趣的簇可视化；DVET（Data Visualization Environment Tool）系统利用虚拟现实技术展示数据空间和空间上的点（数据）；Table Lens系统虽然仍以表的形式表现数据，但是它以图示法代替了表中的数字，并且给出观察的视点，易于用户选择和操纵数据表中的区域；Xgobi系统强调动态和交互技术，用户能同时以不同的可视化方法处理相同的数据。另外，还有更多新的方法和技术正在被研究和开发，如3D技术、基于图形技术等。

综上所述，为了使读者更加详细地了解不同数据可视化技术的特性，表5.1对各种可视化技术从数据特性、任务特性和技术特性三方面做出了对比。

表5.1 可视化技术比较图

可视化技术分类 \ 测试项		聚类	多元分析	变量个数	数据项个数	分类数据	可视化重叠	学习曲线
几何技术	散点矩阵	优	优	良	良	较差	中	优
	山地图	良	良	较差	中	中	良	良
	投影截面组合视图	优	优	良	良	较差	中	良
	多维切片	良	良	良	良	较差	中	中
	平行坐标	中	优	优	较差	中	差	中

（续表）

可视化技术分类 \ 测试项		聚类	多元分析	变量个数	数据项个数	分类数据	可视化重叠	学习曲线
基于图标技术	枝形图	中	中	良	较差	较差	较差	中
	形状编码	中	较差	优	良	较差	良	较差
	彩色图标	中	较差	优	良	较差	良	较差
面向像素技术	独立于查询	良	良	优	优	较差	优	良
	基于查询	良	良	优	优	较差	优	较差
分层技术	多维堆积图	良	良	中	中	优	中	中
	树图	良	中		良	优		中
	锥型树	良	良	中	良	中	良	
基于图形技术	基本图形	中	中	较差	良	中	中	良
	特殊图形	优	良	较差	良	中	良	良

5.3　常用分析工具介绍

“工欲善其事，必先利其器。”要学习数据分析，先要掌握数据分析的基本软件。前面讲到了一些跟统计相关的数据分析方法，除此之外涉及到的高级数据分析都与数据挖掘相关，最后通过可视化技术来展现数据。本节把常用的分析工具分为三大类来介绍，主要包括Excel、SPSS、SAS、Eviews、MATLAB以及前面提到的R语言等；数据挖掘工具主要包括Weka、QUEST、MineSet、Clementine、Rapidminer、Mahout、Darwin、Enterprise Miner、Intelligent Miner、DBMiner等，通过这些工具可以在数据集上进行深度分析；可视化设计工具主要包括D3、Protovis、DataWatch、Tableau、Google Chart等。这些工具仍在迅速发展，不断更新优化，本节主要介绍这三大类中典型的工具。

5.3.1　统计分析工具

1. Excel

Microsoft Excel是微软公司研发的办公软件Microsoft Office的组件之一，是Microsoft为Windows和Apple Macintosh环境编写的一款电子表格软件。Excel是微软办公套装软件的一个重要的组成部分，它可以进行各种数据的处理、分析并用于辅助决策，广泛地应用于管理和统计及财经、金融、零售等众多领域。Excel是一个非常容易入门的软件。使用Excel进行数据分析，简单的分析运用里面最基础的运算和图表的制作就可以了，稍微复杂一点的分析工作可能用到函数和数据透视表，VBA和宏是其数据分析的高级应用。或者可以下载XLstat插件（一个统计分析插件），可以完成大部分SPSS数据分析功能。

在使用分析工具之前，首先得安装Excel的数据分析功能，默认情况下，Excel是不安装这个扩展功能的，安装步骤如下。

（1）鼠标指向Office按钮，然后单击“Excel选项”。

（2）在Excel选项对话框中找到“加载项”，在管理列表框中选择“Excel加载项”，然后单击“转到”。

（3）选择“分析工具库”和“分析工具库–VBA”，单击“确定”。

安装完成后，就可以“数据”板块看到“数据分析”功能了。

2. SPSS

SPSS由美国斯坦福大学的三位研究生Norman H. Nie、C. Hadlai（Tex）Hull 和 Dale H. Bent于1968年成功的开发，它的全称是“Statistical Product and Service Solutions”，中文名为“统计产品与服务解决方案”，是世界上最早的统计分析软件，同时成立了SPSS公司。2009年7月28日，IBM公司宣布用12亿美元现金收购统计分析软件提供商SPSS公司。最初SPSS软件的全称为“社会科学统计软件包”（Solutions Statistical Package for the Social Sciences），但是随着SPSS产品服务领域的扩展和服务深度的增加，SPSS公司已于2000年正式将全称更改为“统计产品与服务解决方案”。这是IBM公司推出的一系列用于统计学分析运算、数据挖掘、预测分析和决策支持任务的软件产品及相关服务的总称。目前有Windows和Mac OS X等版本，标志着SPSS的战略方向做出了重大调整。

SPSS的数据管理和输入方法与Excel很相似，数据接口基本通用，可以很方便地从数据库中读取数据。其内含模型包括常用的、较为成熟的统计分析模型，完全可以满足非统计专业人士的工作需要。另外，SPSS的输出结果十分直观、漂亮，很多都是以图表的形式输出，存储时则使用SPO格式，且可以转存为HTML格式和文本格式。对于熟悉编程运行方式的用户，可直接使用语句生成窗口，只需要在菜单中选择好各个选项，然后粘贴就可以自动生成标准的SPSS程序。SPSS可以直接读取Excel及DBF数据文件，且它的分析结果直观、清晰、易学易用。现已推广到多种操作系统的计算机上，极大地方便了中、高级用户。它成为国际上最有影响的三大统计软件之一①。

SPSS的基本功能包括数据管理、统计分析、图表分析以及输出管理等。统计分析的内容包括描述性统计、均值比较、参数检验、方差分析、非参数检验、一般线性模型、相关分析、回归分析、对数线性模型、聚类分析、生存分析、数据简化、时间序列分析、多维尺度分析、多重响应等多个大类。每类中又有多种专项的统计方法，例如回归分析中又分线性回归分析、非线性回归、曲线估计、Logistic回归、Probit回归、加权估计、两阶段最小二乘法等多个统计过程，而且每个过程中又允许用户选择不同的方法及参数。对于分析结果的展现，SPSS有专门的绘图系统，可以根据数据和用户的要求绘制各种图形。

SPSS提供了三种运行方式：完全窗口菜单运行方式、程序运行方式和批处理方式。完全窗口菜单方式非常适合非专业人士使用，不需要编程就可以使用。程序运行方式和批处理方式则是从使用者特殊的分析出发，需要掌握SPSS编程语法，对使用者的要求较高。目前，SPSS在我国的社会科学、自然科学的各个领域发挥了巨大作用。该软件还可以应用于经济学、数学、统计学、物流管理、医学、生物学、心理学、地理学、体育、农业、林业、市场营销等领域。

3. SAS

SAS最早由北卡罗来纳大学的两位生物统计学研究生开发，系统全称为“Statistics

① http://baike.baidu.com/view/130328.htm?fromtaglist.

Analysis System”，于1976年正式推出。SAS主要用于大型集成信息系统的决策支持，最初它的功能仅限于统计分析，至今，它的重要组成部分和核心功能也仍是统计分析功能。全世界接近三万家机构采用SAS软件进行分析，它的直接用户已经超过三百万人，被称为统计软件界的巨无霸。

SAS系统是由多个功能模块组合而成的组合软件系统，其中BASE SAS模块是SAS系统最基本的、最核心的部分。它的功能很强大，主要负责数据管理任务，并管理用户使用环境，处理用户语言，还是SAS系统的中央调度室，调用其他SAS模块和产品。它不仅能够单独存在，也能够与其他产品或模块共同构成一个完整的系统。

SAS系统具有灵活的功能扩展接口和强大的功能模块，在BASE SAS的基础上，还可以通过增加不同的模块来增加不同的功能，如SAS/STAT（统计分析模块）、SAS/GRAPH（绘图模块）、SAS/QC（质量控制模块）、SAS/OR（运筹学模块）、SAS/ETS（经济计量学和时间序列分析模块）、SAS/IML（交互式矩阵程序设计语言模块）、SAS/AF（交互式全屏幕软件应用系统模块）SAS/FSP（快速数据处理的交互式菜单系统模块）等。在数据结果展现方面，SAS也有一个智能型绘图系统，不仅能绘各种统计图，还能绘出地图。SAS提供多个统计程序，每个程序均含有多种选项，用户还可以通过对数据集进行加工，实现更为复杂的统计分析。此外，SAS还提供各类概率分析函数、分位数函数、样本统计函数和随机数生成函数，方便用户实现特殊的统计要求①。

4. Eviews

Eviews是Econometrics Views的缩写，直译为计量经济学观察，通常称为计量经济学软件包。它主要是采用计量经济学方法与技术对社会经济关系与经济活动的数量规律进行“观察”。计量经济学研究的核心是收集资料、设计模型、估计模型、检验模型、应用模型（结构分析、经济预测、政策评价）。该软件是专门运行在Windows环境下从事数据分析、回归分析和预测的工具。利用Eviews可以迅速地从数据中寻找出统计关系，并用得到的关系去预测数据的未来值②。

Eviews是计量经济学软件中的世界性领导软件。Eviews采用了革新的图表界面和准确的分析引擎工具，使得EViews具有强大、灵活和容易使用的功能。Eviews的应用范围包括：宏观经济预测、金融分析、科学实验数据分析与评估、销售预测、仿真和成本分析等。

5. MATLAB

MATLAB是matrix&laboratory两个词的组合，意为矩阵实验室（矩阵工厂）。是由美国MathWorks公司出品的商业数学软件，主要包括MATLAB和Simulink两大部分。主要面向科学计算、可视化以及交互式程序设计的高科技计算环境，被用于算法开发、数据可视化、数据分析以及数值计算等场合。

6. Stata

Stata是Statacorp于1985年开发的一套用于数据分析、数据管理以及绘制专业图表的完整

① http://wenku.baidu.com/view/d3db732ccfc789eb172dc8e7.html.

② http://baike.baidu.com/view/207806.htm?fr=aladdin.

及整合性的软件。它提供了一系列的功能，包含线性混合模型、均衡重复反复及多项式普罗比模式。

最新版本的Stata采用最具亲和力的窗口接口，当用户自行编写程序时，软件可以提供具有直接命令式的语法。另外，Stata还提供完整的使用手册，包含样本的建立、解释、模型与语法、文献等超过一万余页的文档。

Stata的功能模块包括三个部分。

- 统计功能：Stata的统计功能很强，除了传统的统计分析方法外，还包括近20年发展起来的新方法，如指数与Weibull回归，Cox比例风险回归，多类结果与有序结果的logistic回归，Poisson回归，负二项回归及广义负二项回归，随机效应模型及矩阵运算等。
- 作图功能：主要提供八种基本图形的制作，包括直方图（histogram）、条形图（bar）、百分条图（oneway）、百分圆图（pie）、散点图（twoway）、散点图矩阵（matrix）、星形图（star）和分位数图。对这些图形的巧妙应用，能够满足绝大多数用户作图的要求。在一些非绘图命令中，还提供了专门绘制某种图形的功能，例如，在生存分析中提供了绘制生存曲线图，在回归分析中提供了残差图等。
- 程序设计功能：Stata不仅是一个统计分析软件，而且为用户二次开发提供了很强的程序设计功能。这为用户提供了一个广阔的开发应用的天地，用户可以充分发挥自己的才能，熟练应用各种技巧，真正做到随心所欲地完成需要的工作。事实上，Stata的ado文件（高级统计部分）就是采用Stata提供的语言编写的。

Stata的分析能力很强大，甚至在许多方面远远超过了SPSS和SAS。Stata计算速度极快，因为它在分析时是将数据全部读入内存而不是磁盘中，在计算全部完成后才和磁盘交换数据。Stata也可采用命令行方式来操作，但使用上比SAS简单很多。其生存数据分析、纵向数据分析等模块的功能很出色，大大超过了SAS。另外，用Stata绘制的分析结果图形相当精美，很有特色。Stata在全球范围内被广泛应用于企业和学术机构中，许多使用者是工作在特定研究领域一线的人员，比如经济学、社会学、政治学及流行病学等领域[①]。

5.3.2 数据挖掘工具

1. Weka

Weka（Waikato Environment for Knowledge Analysis）的全名是怀卡托智能分析环境，Weka的主要开发者来自新西兰的一所大学The University of Waikato。它是一款免费的，非商业化的，基于Java环境下开源的机器学习以及数据挖掘软件。Weka系统得到了广泛的认可，被誉为数据挖掘和机器学习历史上的里程碑，是现今最完备的数据挖掘工具之一（已有11年的发展历史），每月下载次数已逾万次。

Weka作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，高级用户可以通过Java编程和命令行来调用其分析组件。同时，Weka也为普通用户提供了图形化界面，称为Weka KnowledgeFlow Environment和Weka Explorer。Weka与R语言相比，虽然在统计分析方面较弱，但在机器学习方面要强得多，它包括对数据进行预处理，分

① http://baike.baidu.com/view/1141894.htm?fr=aladdin.

类，回归、聚类、关联规则以及在新的交互式界面上的可视化。可以在Weka论坛里（http://weka.sourceforge.net/wiki/index.php/Related_Projects）找到很多扩展包，比如文本挖掘、可视化、网格计算等。很多其他开源数据挖掘软件也支持调用Weka的分析功能。

2. QUEST

QUEST[①]是一个多任务数据挖掘系统，由IBM公司Almaden研究中心开发。该系统提供了高效的数据开采基本构件可用于新一代决策支持系统的应用开发。

QUEST具有以下特点。

- 具有查全性的算法，能找出所有满足指定类型的模式。
- 各种开采算法都具有近似线性（O（n））的计算复杂度，可适用于任意大小的数据库。
- 为各种发现功能设计了相应的并行算法。
- 提供了专门在大型数据库上进行各种开采的功能，包括序列模式发现、关联规则发现、决策树分类、时间序列聚类、递增式主动开采等。

3. MineSet

MineSet是一个多任务数据挖掘系统，由Standford大学和SGI公司联合开发。它将可视化工具和多种数据挖掘算法结合起来，为用户在挖掘、理解隐藏在数据背后的大量知识或规律时提供更加实时、直观的帮助。

MineSet系统具有如下特点。

- 先进的可视化展现方法。为使用户能够用不同的可视化工具对同一个挖掘结果以多种形式表示，MineSet中使用了6种可视化工具来表现数据和知识。
- 多种数据转换功能。在进行数据挖掘前，可对不必要的数据项进行剔除，可对数据进行统计、集合、分组，转换数据类型，构造表达式，然后由已有的数据项生成新的数据项，以及对数据进行采样等。
- 多种数据挖掘模式。主要包括关联规则、分类器、聚类归类、判断列重要度、回归模式等。
- 支持多种关系数据库。可以借助SQL命令执行查询，也可直接从Oracle、Informix、Sybase的表中读取数据。
- 支持国际字符。
- 可直接发布到Web。
- 操作简单。

4. Clementine

Clementine是SPSS发行的一种数据挖掘工具，也称为“IBM SPSS Modeler”。最早是由ISL（Intergral Solutions Ltd.）公司开发的一款数据挖掘产品，后来该公司被SPSS公司收购，之后SPSS公司对该产品进行一系列的技术改造和优化。Clementine是采用客户/服务器架构的产品，既可以单机运行，也能够连接到网络上的Clementine Server。此工具结合了Neural Networks、Association Rules及Rule-induction Techniques等具多种图形使用接口的分析技术，

① http://wenku.baidu.com/view/2a8b7fb069dc5022aaea0074.html.

使得Clementine不仅具有分析功能，还能够提供可使用的简单的可视化程序环境，Clementine的数据可视化能力包括散布图、平面图及Web分析。Clementine包含flat file以及关系型数据库（经由ODBC），资料存取能力强大，而且为用户提供了大量的人工智能、统计分析的模型（神经网络，聚类分析、关联分析、因子分析等），并用基于图形化的界面为用户认识、了解、熟悉这个软件提供了极大的方便。此外，它能够与SPSS统计功能有更多的整合，数据处理也更加灵活好用。

Clementine的设计思想是用简单的方式进行数据挖掘，尽量屏蔽数据挖掘算法的复杂性以及软件操作的繁琐性，使数据挖掘人员将更多的精力放在使用先进的挖掘技术解决商业问题而不是操作软件本身。

作为一个数据挖掘平台，Clementine结合商业经验能够快速建立预测性模型，进而应用到商业活动中，帮助用户改进决策过程。与其他数据挖掘工具相比，Clementine强大功能的数据挖掘算法，使数据挖掘贯穿业务流程的始终，在缩短投资回报周期的同时极大地提高了投资回报率，而不是仅注重模型的外在表现，忽略数据挖掘在整个业务流程中的应用价值。

5. Rapidminer

RapidMiner是世界上领先的数据挖掘解决方案，在非常大的程度上有着先进技术。它的数据挖掘任务涉及范围广泛，包括各种数据艺术，能简化数据挖掘过程的设计和评价。RapidMiner以前称为YALE，它提供了图形化界面，采用类似Windows资源管理器中的树状结构来组织分析组件，树上每个节点表示不同的运算符。它是基于Java开发，用Weka来构建的，即它可以调用Weka中的各种分析组件。耶鲁大学已成功地将其应用在各种领域，主要包括数据流挖掘，文本挖掘，多媒体挖掘，分布式数据挖掘和集成开发。

Rapidminer具有如下特点。

- 免费提供数据挖掘技术和库。
- 全部采用Java代码（可运行在大部分操作系统上）。
- 数据挖掘过程简单，强大和直观。
- 内部XML保证了标准化的格式来表示交换数据挖掘过程。
- 可以用简单脚本语言自动进行大规模进程调用。
- 多层次的数据视图，确保数据的有效和透明。
- 图形用户界面的互动原型。
- 命令行（批处理模式）自动进行大规模应用。
- 支持Java API（应用编程接口）。
- 简单的插件和推广机制。
- 强大的可视化引擎，提供了许多尖端的高维数据的可视化建模。
- 拥有400多个数据挖掘运营商支持。

6. DBMiner

DBMiner[1]以前称作DBLearn，它是一个多任务数据挖掘系统，由加拿大Simon Fraser大学

① http://baike.baidu.com/view/3079008.htm.

开发。DBMiner系统具有如下特点。

- 将多种数据挖掘技术综合。如统计分析、面向属性的归纳、元规则引导发现、逐级深化发现多级规则等技术。
- 完成多种知识的发现：如泛化规则、关联规则、特性规则、分类规则、演化和偏离知识等。
- 数据挖掘查询语言DMQL，它是一种交互式的类SQL语言。
- 实现了基于客户/服务器体系结构的Unix和PC（Windows/NT）版本的系统。
- 平滑集成关系数据库。

DBMiner系统的设计特点是集成数据开采和关系数据库，然后借助于面向属性的多级概念来发现各种知识。

7. Mahout

Mahout是Apache Software Foundation（ASF）旗下的一个开源项目，在机器学习领域提供了一些可扩展的经典算法的实现和数据挖掘的程序库。它可以实现很多功能，包括聚类、分类、推荐过滤、频繁子项挖掘等。该项目很大部分建立在Hadoop分布式计算项目的基础上，把许多以前在单机上运行的算法，转化为MapReduce模式，从而大大提升了算法可处理的数据量和处理性能。此外，通过使用Apache Hadoop库，可以将Mahout有效地扩展到云中。但是它只提供给开发者使用的一个工具框架，并不提供用户接口、预打包的服务器或者安装程序。

Mahout实现的机器学习算法包括：聚类算法、分类算法、关联规则挖掘、回归、降维（维约简）、进化算法、过滤、向量相似度计算、非Map-Reduce算法等。

Mahout支持一些集群算法（都是使用Map-Reduce编写的），这些算法都有一组各自的标准和目标，如Canopy是一种快速集群算法，常用于为其他集群算法创建初始种子；k-Means（以及模糊k-Means）是根据项目与之前迭代的质心（或中心）之间的距离将项目添加到k集群中；Mean-Shift是无需任何关于集群数量的推理知识的算法，可以生成任意形状的集群；Dirichlet借助基于多种概率模型的集群，不需要提前执行特定的集群视图。

5.3.3　可视化设计工具

读者在本章了解了数据分析以及可视化的技术，对于想要实现数据可视化的用户可能会问“有没有什么软件可以实现数据可视化”。很幸运，目前有很多开箱即用的可视化软件，只需有鼠标就能够操作，还有一些软件则需要一点编程技巧，虽然有些工具并不是专门用于制作数据图，但依然有帮助，下面会列举当前比较流行的几种可视化工具。

1. Many Eyes

Many Eyes是IBM视觉传达实验室主导的一个研究项目，该软件带有一系列交互式的可视化工具，也是一个在线应用，能够识别带分隔符的文本文件。Many Eyes名字的由来是因为其初衷是想了解人们能否以群组的形式探索大型数据集。如果一个群组内的众多眼睛来观察某个数据集，是否可以从中挖掘到更多有意思的地方？效率是否会更高？

虽然Many Eyes目前尚未提供多人的数据分析功能，但作为个人使用来说它依然很有价

值。该软件提供的功能涵盖了绝大多数传统的可视化类型，它的优势在于其中的可视化数据图都是可交互的，而且有一些定制选项。比如说在散点图中，可以用第三种指标来测量各个数据节点，而且鼠标悬停在感兴趣的数据节点时还能查看具体的数值。

Many Eyes是目前数据探索中用途最为广泛的免费工具之一，不过在使用该软件时有两点仍然值得注意。其一，该软件的大部分工具都是Java应用小程序，在使用的电脑中如果没有安装Java，可能就无法充分利用它；其二，对多数人来说可能更为敏感，因为使用该软件时上传到网站的数据是存储在公共空间里的，所以最好不要用Many Eyes来挖掘贵公司的客户信息或者销售数据。

2. D3

D3是纽约时报可视化编辑Mike Bostock与他斯坦福的教授和同学合作开发的用于数据文件处理的JavaScript Library，全称叫做Data-Driven Document，Mike Bostock也是Protovis的开发者之一。D3的最大特性就是能把数据和文档对象模型（DOM）结合，从而对文档进行数据驱动的操作和交互。例如，用户可以用D3从数组生成HTML表格，或者使用相同数据平滑和动态规律创建一个SVG图表。D3的轻量级特性使它能够更好地利用CSS 3、HTML 5和SVG（Scalable Vector Graphics，可伸缩向量图形）等底层技术。D3的性能非常出色，速度极快，支持大数据集和动态交互及动画效果，它可以非常灵活地设计出Web可视化应用。此外，它既可以作为一个可视化框架（如Protovis），也可以作为构建页面的框架（如jQuery）。D3的功能设计允许代码重用，通过集合不同的组件和插件，对视图结果有很大的可控性。它支持的数据格式常见的有：txt、html、json、html、xml、csv等。

D3是一款开源的可视化工具，它的优点是动画和交互图，入手简单，易于使用，降低了服务器负载，拥有大量的图表类型，跨服务器语言。但是可能不符合我们中国人的报表使用习惯。

3. Protovis

Protovis是一款免费的开源JavaScript可视化工具，由斯坦福可视化小组的Mike Bostock和Jeff Heer开发，该软件主要的功能是从用户角度观察数据，并用简单的图形描述数据。Protovis通过动态地继承、扩展、布局避免了图形库随着功能的增强而变得臃肿这个缺陷，这是许多可视化图形库达不到的。它用JavaScript和SVG构建Web可视化应用，无需插件。Protovis也可通过示例教学进行学习。

4. Datawatch

Datawatch是Datawatch公司开发的数据可视化工具，它可以直接获取结构化和非结构化数据，并以可视化的方式实时展现数据，它能与市面上所有的数据源实现无缝对接，包括DATABASE、KDB+和OData等。同时可以实现在数据运转中实时查看数据，更直观地将稠密数据可视化，更立体地分组显示数据，更流畅地显示快速变化的数据。该软件可以快速地完成ETL，灵活地进行数据转化，有效获取半结构化的数据，兼容现有的方法以及用SQL语言获取数据。当数据与预设值产生偏差时，可以提供警告服务，并支持嵌套与扩展。主要产品包括：Datawatch Desktop和Datawatch Server。它可应用于数据庞大且繁琐、实时要求性较高的需

求，如销售业绩、风控等。

Datawatch支持多种信息可视化技术，包括：树形图、热力图、矩阵热点图、柱形图、地平线图、线形图、针状图、OHLC（开盘价、最高价、最低价、收盘价）图、散点图、堆栈图、表面可视化图。Datawatch可以支持以上可视化图表以及其他更多图表，能帮助用户看到多种层级结构、内在关系以及各个对象的细节，从而解决问题、了解复杂的关系，找出需要关注的领域。

5. Quadrigram

Quadrigram是西班牙的Bestiario开发的可视化编程语言，致力于收集数据、处理数据、分享数据、可视化信息。前身叫做“Impure”。它可从各种数据中提取信息，无论是用户自己的数据还是来自社交媒体的数据、财经数据、图片、新闻、搜索结果等。利用模块化的逻辑接口，用户可以快速地设置交互方式，借助于可视化方法呈现数据。灵活与直观地建立用户自定义的数据可视化视觉语言，它能够使用户快速标准化并分享自己的想法，分析和监控用户的数据，以及以交互式可视化的形式，通过动画或指示板产生令人信服的解决方案。

6. Tableau

Tableau是桌面系统中最简单的商业智能工具软件，是一款只面向Windows系统的软件，它的设计初衷主要是用于视觉化的数据研究和分析，其最大的优势是在美学和设计上比其他软件要突出。它没有强迫用户编写自定义代码，新的控制台也可完全自定义配置。在控制台上，不但可以监测信息，而且还提供完整的分析能力。Tableau控制台不仅具有灵活性，而且具有高度的动态性。Tableau发布了免费版本的Tableau Public。

Tableau公司完美地将数据运算与美观的图表嫁接在一起。它的程序很容易上手，用户可以直接用它将大量数据拖放到数字“画布”上，快速地创建好各种图表。这一软件的理念是，界面上的数据越容易操控，公司就能对自己所在业务领域里的所作所为是否正确了解得越透彻。

Tableau Desktop是基于斯坦福大学研发的突破性技术的应用程序。它帮助用户生动地分析实际存在的任何结构化数据，并在几分钟内生成美观的图表、坐标图、仪表盘与报告。利用Tableau简便的拖放式界面，用户可以自定义视图、布局、形状、颜色等，帮助用户展现自己的数据视角。Tableau Server也是应用程序，它可以将Tableau Desktop中最新的交互式数据可视化内容、仪表盘、报告与工作簿的共享变得简单快速。Tableau Reader是免费的计算机应用程序，帮助用户查看内置于Tableau Desktop的分析视角与可视化内容。拥有Tableau Interactor许可证的用户可以交互、过滤、排序与自定义视图。如果是拥有Tableau Viewer许可证的用户则可以查看与监视发布的视图。

7. Google Chart

Google Chart提供了一种非常完美的方式来可视化数据，它属于在线工具。它提供了大量现成的图表类型，从简单的线图表到复杂的分层树、地图等都有。它还内嵌了动画和用户交互控制。

除此之外，Google还提供了一系列免费的可视化组件，下面介绍它的几个主要组件。（注：Google在线应用正在不断更新。）

Google Chart API让用户能通过URL传递参数，生成动态的图表和图片。该API可以产生各种各样的图表，如饼图、地图、QR码和文氏图等。所有描述图片的参数都包含在URL中。部分图表的URL可以采用Chart Wizard快捷地生成，生成的URL可以嵌入<img>标签中。

Google Visualization API是对Google Chart API的补充和提升，它可以用来开发更高级的网络版的交互图表和图片，可以直接从Google Docs平台等在线数据库中获取图表的数据。Google Visualization API生成的图表有丰富的交互性，用户可以直接编码来处理事件，实现更好的网页效果。用户可以编写JavaScript和HTML或者通过Google Gadget的小工具来设计自己的想要的报告和交互界面，可视化地分析、显示数据。Google Visualization API也可以让用户创建、分享和重用开发者社区构建的可视化工具。

Google Ngram Viewer可以查询并可视化某个单词或词组在过去500年的书中出现的频率。基于Google富于争议而雄心勃勃的图书数字化计划获取的海量数据，Ngram Viewer引擎可以分析词汇在海量图书中使用的频率和使用概率的历史变化趋势。

Google Analytics是用来分析网站流量和用户行为的工具。该工具会跟踪并且记录用户在网站上的访问行为、行为统计、用户地理位置等各种信息，最终的结果以可视化的形式呈现在交互界面上，这些可视化的形式包括条状图、折线图、火花线、饼图、运动图和区域地图等。

8. JFreeChart

JFreeChart是基于Java平台的一个开放式的图表绘制类库。它完全使用Java语言编写，是为Applications、Applets、Servlets以及JSP等使用所设计的。JFreeChart可生成多种图表，如饼图、柱状图、散点图、时序图、甘特图等，并且可以生成PNG和JPEG格式的图片，还可以与PDF和EXCEL关联。

5.4 练习

一、选择题

1. Hive查询语言和SQL的不同之处在于______操作。

A. Group By　　B. Join

C. Partition　　D. Union

2. Hive最终重视的性能是可测量性，延展性，______和对于输入格式的宽松匹配性。

A. 较低恢复性　　B. 容错性

C. 快速查询　　D. 可处理大量数据

3. 下面哪些工具不属于可视化工具？______

A. Many Eyes　　B. Datawatch

C. Mahout　　D. Tableau

二、简答题

1. Hive的体系结构具体包括什么？

2. Stinger是对Hive的优化的项目，具体的改进包括什么？

3. 可视化技术有哪几类？分别是什么？
4. 时间序列分析中怎么建立模型？
5. 主成分分析与因子分析的异同点？
6. 数据分析的定义？
7. 可视化的定义？

参考文献

[1] Nathan Yau.《鲜活的数据——数据可视化指南》[M]. 北京：人民邮电出版社，2012.
[2] 张文霖，刘夏璐，狄松.《谁说菜鸟不会数据分析》[M]. 北京：电子工业出版社，2011.
[3] 赵刚.《大数据技术与应用实践指南》[M]. 北京：电子工业出版社，2013.
[4] 韩家炜.《数据挖掘概念与技术》[M]. 北京：机械工业出版社，2007.
[5] Julie Steele，Noah lliinsky.《数据可视化之美》[M]. 北京：机械工业出版社，2011.
[6] 刘堪，周晓峥. 数据可视化研究与发展[J]. 计算机工程，2002.8（28）.
[7] 张文彤，钟云飞. IBM SPSS数据分析与挖掘实战案例精粹[M]. 北京：清华大学出版社，2013.
[8] 任永功，于戈. 数据可视化技术的研究与进展[J]. 计算机科学，2004.12.
[9] 李佳书. 多元图表示原理的可视化分类方法研究[D]. 知网，2006.
[10] 刘圆圆. 时间序列分析及其应用[J]. 科技创新导报，2011（27）：255.
[11] 王燕. 应用时间序列分析（第二版）[M]. 北京：中国人民大学出版社，2008.
[12] 杨桦. 基于Excel的最优子集多元线性回归预测模型设计[J]. 中国管理信息化，2011，14（18）：70–71.

第6章 数据挖掘技术

数据采集和数据存储技术的不断进步使组织积累了海量的数据，而且这些数据量还在不断地快速增长。快速增长的海量数据存储在数据库、数据仓库中，从中提取有用信息成为了巨大的挑战。早在1982年，趋势大师约翰·奈斯比（John Naisbitt）就在他的首部著作《大趋势》（Megatrends）中提到“人类正被信息淹没，却饥渴于知识。”由于数据量太大，并且数据本身具有新的特点，传统的数据分析工具和技术已无法完全满足海量信息处理的需求。此时，数据挖掘技术则将传统的统计分析方法与处理大量数据的复杂算法结合起来，为探查和分析新的数据类型以及用新方法分析海量数据提供了契机。

6.1 数据挖掘简介

数据挖掘（Data Mining，DM）简单来说就是在大量的数据中提取或挖掘信息，通过仔细分析来揭示数据之间有意义的联系、趋势和模式。数据挖掘技术出现于20世纪80年代后期，是数据库研究中的一个新领域，具有很高的研究与应用价值。数据挖掘属于交叉性学科，它融合了统计学、数据库技术、机器学习、人工智能、模式识别和数据可视化等多个领域的理论和技术，如图6.1所示。

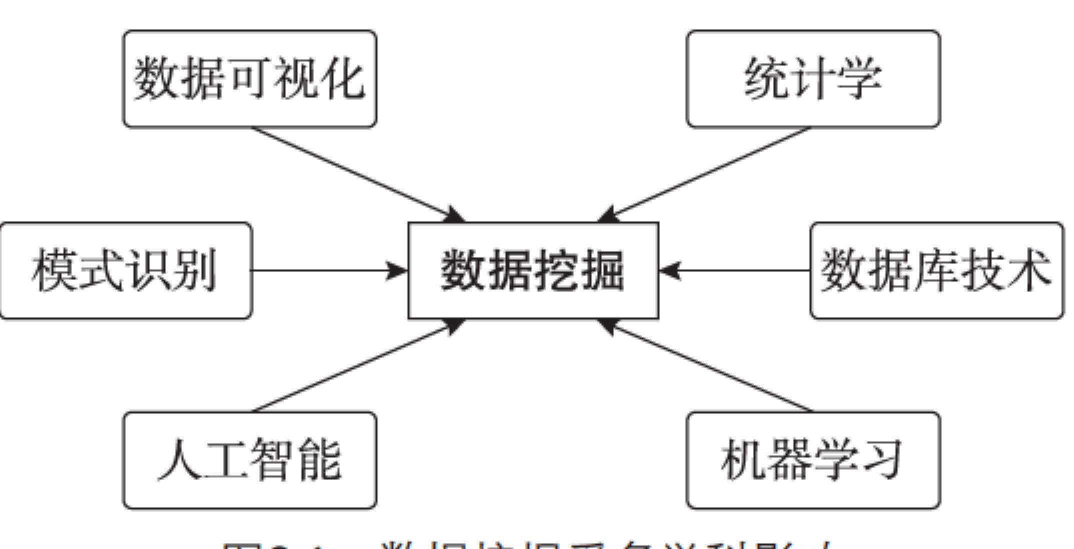


图6.1 数据挖掘受多学科影响

数据挖掘主要是为了发现隐藏在数据中的有用信息和规律，数据库中知识发现（Knowledge Discovery in Database，KDD）是将未加工的数据转换成有用信息的整个过程，因而从模式处理的角度，许多人对这二者并没有作严格的区分，但本书认为数据挖掘只是KDD中的一个核心步骤，如图6.2所示。

（1）数据清理：消除噪声或不一致的数据。

（2）数据集成：将多种数据源集合到一起。

（3）数据转换和选择：提取相关数据，并将其转换成适合挖掘的形式。

（4）数据挖掘：KDD的关键步骤，运用相关的数据挖掘技术得到数据模式。

（5）模式评估：根据兴趣度度量，识别表示知识的有用的模式。

（6）知识表示：运用知识表示、可视化等相关技术，向用户展示挖掘到的知识。

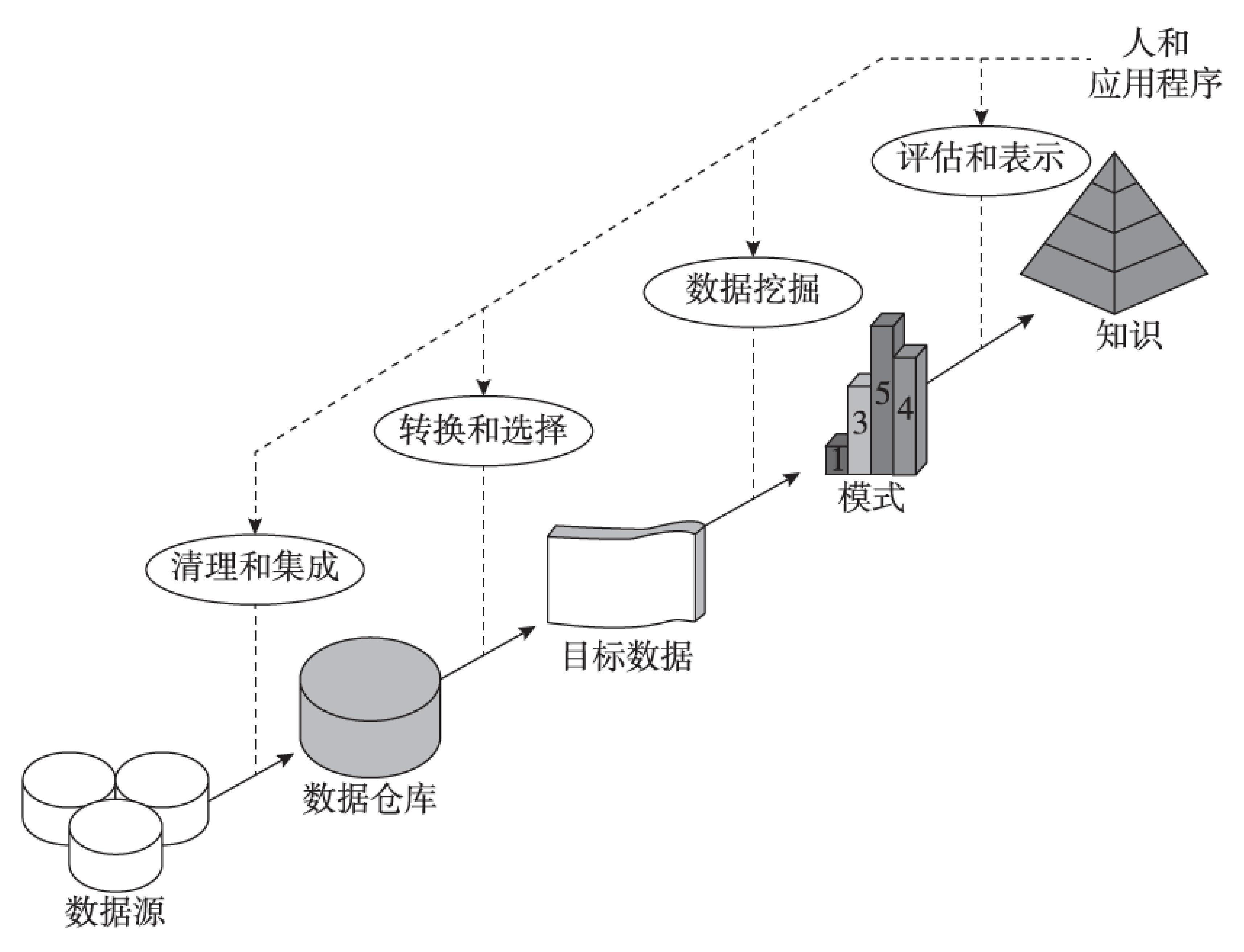
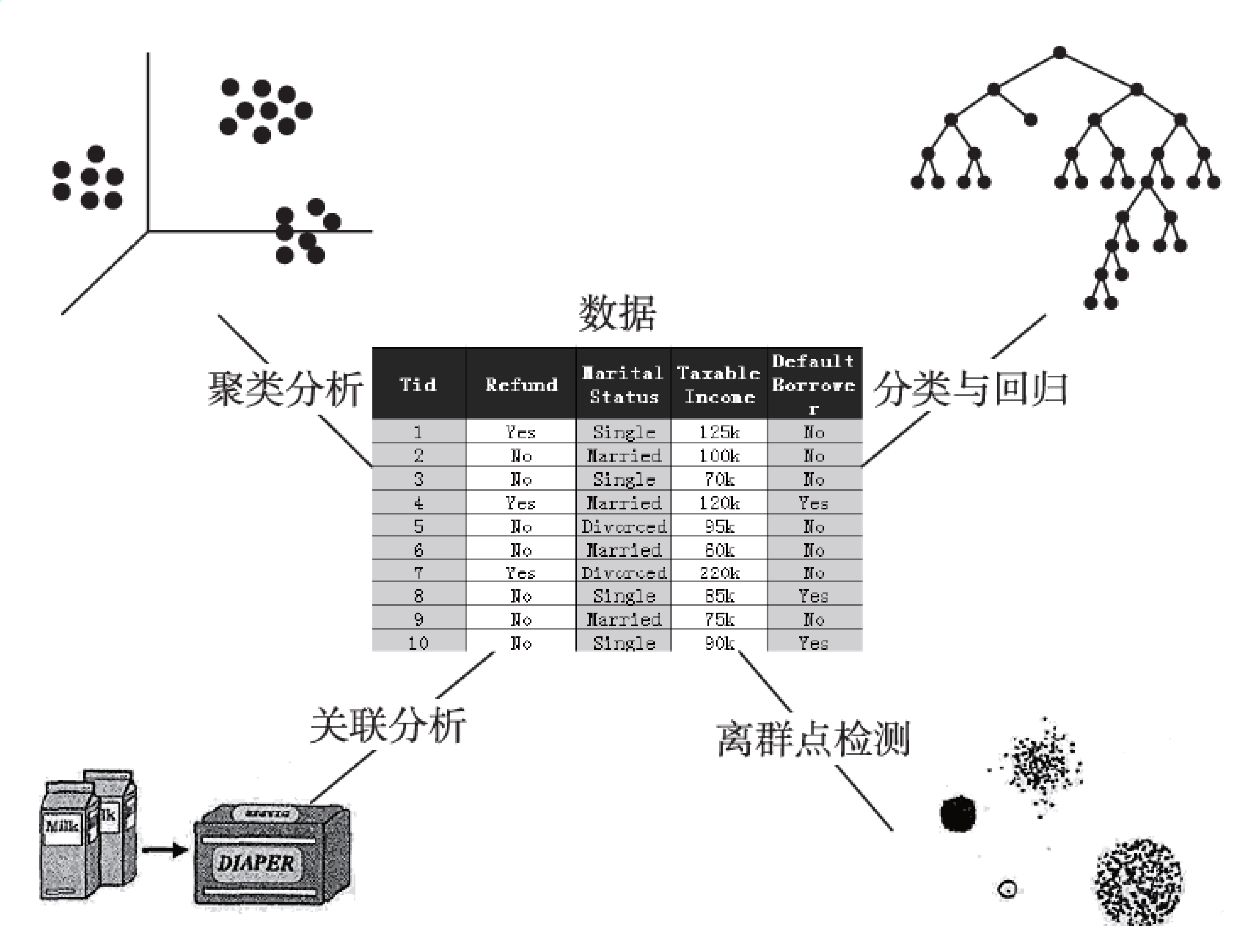


图6.2 数据库中知识发现的过程

一般地，数据挖掘任务可以分为两大类。

- 描述任务：刻画数据的特征，概括数据中潜在联系的模式（包括相关，趋势，聚类和异常等）。
- 预测任务：根据当前数据进行推理、预测，根据其他属性的值，预测特定属性的值。

本质上，描述性挖掘任务通常是探查性的，并且常常需要后处理技术来验证和解释结果。图6.3展示了本章介绍的四种主要数据挖掘任务，包括关联分析、分类与回归、聚类分析和离群点检测。



Tid	Refund	Marital Status	Taxable Income	Default Borrower
1	Yes	Single	125k	No
2	No	Married	100k	No
3	No	Single	70k	No
4	Yes	Married	120k	Yes
5	No	Divorced	95k	No
6	No	Married	60k	No
7	Yes	Divorced	220k	No
8	No	Single	85k	Yes
9	No	Married	75k	No
10	No	Single	90k	Yes

图6.3 四种主要数据挖掘类型

数据挖掘产生于应用，面向于应用。数据挖掘的跨行业数据挖掘标准过程（Cross Industry Standard Process for Data Mining，CRISP-DM）是当今数据挖掘界通用的标准之一，它强调数据挖掘在商业中的应用，解决商业中的问题。CRISP-DM参考模型中包括：商业理解、数据理解、数据准备、建立模型、模型评估和模型部署六个阶段，如图6.4所示。

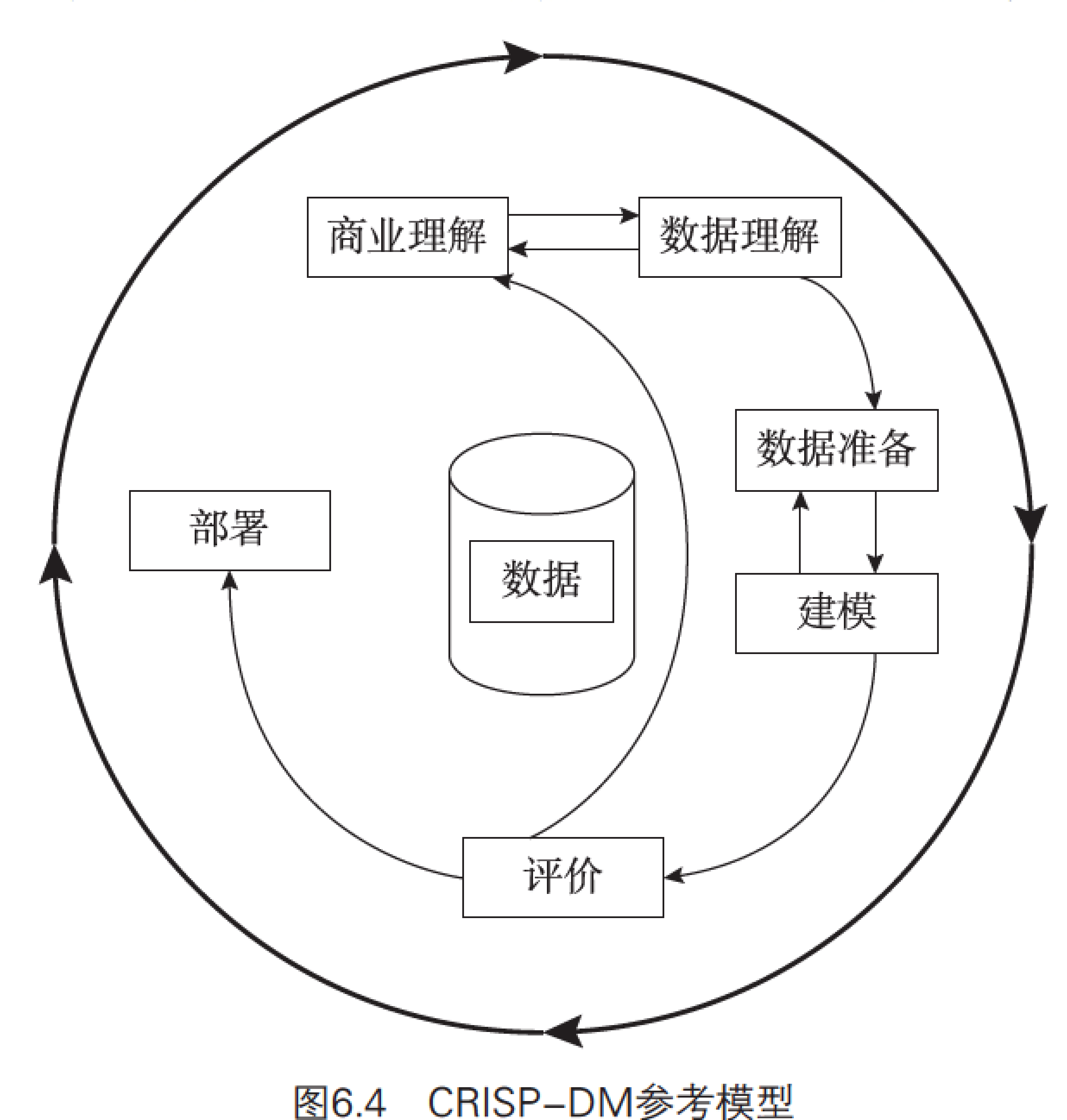


图6.4　CRISP-DM参考模型

（1）商业理解：从商业角度理解项目的目标和要求，把这些理解转换成数据挖掘问题的定义和实现目标的最初规划。

（2）数据理解：从收集数据开始，然后熟悉数据、甄别数据质量问题、发现对数据的真知灼见或探索出令人感兴趣的数据子集并形成对隐藏信息的假设。

（3）数据准备：为建模工具准备适合挖掘的数据类型的过程，涵盖从原始数据到最终数据集的全部活动，主要包括数据的转换和清洗。

（4）建立模型：运用适当的建模技术，建立科学合理的模型，并优化模型中的参数。

（5）模型评估：对所建模型进行较为全面的评价，重审建立模型的步骤，并确认模型能否达到商业目的，从而不断地调整并优化模型，并确定使用数据挖掘结果得到的决策是什么。

（6）模型部署：生成报告或者实施一个覆盖企业的可复用的数据挖掘过程。

6.2　关联分析

关联分析就是从有噪声的、模糊的、随机的海量数据中，挖掘出隐藏的、人们事先不知道、但是有潜在关联的信息或知识的过程，所发现的信息或知识通常用关联规则或频繁项集的形式表示。随着收集和存储在数据库中的数据规模的增大，人们对从这些数据中挖掘出的关联知识也越来越有兴趣。例如：从大量的商业交易记录中发现有价值的关联知识就可帮助

企业进行交叉营销、客户关系管理或辅助相关的商业决策。表6.1给出了一个通常称作购物篮事务的例子，表中每行对应一个事务，包含一个惟一标识TID和给定顾客购买的商品的集合。零售商对分析这些数据很感兴趣，这便于他们了解顾客的购买行为，进而采取对应的促销活动。

表6.1　购物篮事务的例子

TID	项　集
1	{面包，牛奶}
2	{面包，尿布，啤酒，鸡蛋}
3	{牛奶，尿布，啤酒，可乐}
4	{面包，牛奶，尿布，啤酒}
5	{面包，牛奶，尿布，可乐}

从表6.1所示的数据中可以提取出如下规则：

{尿布}→{啤酒}

该规则表明尿布和啤酒的销售之间存在着很强的联系，因为许多购买尿布的顾客同时也购买了啤酒。零售商可以利用这类规则，帮助他们发现新的交叉销售商机。除了购物篮分析外，关联分析也可以应用于其他领域，如生物信息学、医疗诊断、网页挖掘和科学分析等。

在对数据进行关联分析时，需要注意两个关键的问题：第一，从大型事务数据集中发现模式在计算上可能要付出很高的代价；第二，所发现的模式有可能是虚假的，因为发现的模式可能是偶然发生的。本节主要围绕这两个问题组织、介绍关联分析的基本概念和算法等。

6.2.1　基本概念

设$I=\{i_1,i_2,\cdots,i_m\}$是项的集合，其中$i_k\{K=1,2,...,m\}$表示项，如果$X\subset I$，集合X被称为项集，如果一个项集包含k个项，则称它为k-项集。事务二元组$T=(TID，X)$，TID是事务惟一的标识符，称为事务号，数据集$D=(t_1,t_2,...,t_n)$是由$t_1,t_2,\cdots,t_n$事务组成的集合。

如果项集X是事务t_j的子集，则称事务t_j包含项集X。项集的一个重要性质是它的支持度计数，即包含特定项集的事务个数，项集X的支持度计数$\delta(X)$可以表示为：

$$\delta(X)=\left|\{t_i \middle| X\leqslant t_i,\ t_i\in T\}\right| \tag{6-1}$$

其中，符号|·|表示集合中元素的个数。

关联规则可以描述为$A\Rightarrow B$的蕴含式，其中，$A\subset I, B\subset I$，并且$A\cap B\neq\varnothing$。关联规则的强度可以用它的支持度（support，s）和置信度（confidence，c）度量，支持度表示D中事务包含$A\cup B$的百分比，它是概率$p(A\cup B)$，而置信度表示包含A的事务也包含B的百分比，它是条件概率$p(B\mid A)$。支持度和置信度这两种度量的形式定义如下：

$$s(A\Rightarrow B)=p(A\cup B)=\frac{\delta(\mathrm{A}\cup\mathrm{B})}{\delta(\mathrm{D})} \tag{6-2}$$

$$\mathrm{c}(A\Rightarrow B)=p(B|A)=\frac{\delta(\mathrm{A}\cup\mathrm{B})}{\delta(\mathrm{A})} \tag{6-3}$$

关联分析是在事务D中找出大于用户所给定的最小支持度（minsup）和最小置信度（minconf）的关联规则。关联规则的挖掘问题通常可以分解为以下两个子问题。

（1）产生频繁项集：发现满足最小支持度阈值的所有项集，这些项集被称为频繁项集。

（2）规则的产生：从上一步发现的频繁项集中提取所有满足所要求的置信度的规则，这些规则称作强规则。

通常，产生频繁项集所需要的计算开销远大于产生规则所需要的计算开销。

6.2.2 经典频集算法

Apriori算法是一种最有影响的挖掘布尔关联规则频繁项集的算法，算法有两个关键步骤：一是发现所有的频繁项集；二是生成强关联规则。

Apriori算法的核心思想如下。

对于给定的一个数据库，首先对其进行扫描，找出所有的频繁1–项集，该集合记做L_1，然后利用L_1找频繁2–项集的集合L_2，L_2找L_3，如此下去，直到不能再找到任何频繁k–项集。最后在所有的频繁集中提取出强规则，即产生用户所感兴趣的关联规则。

Apriori算法扫描数据库的次数等于最大频繁集的项数。Apriori算法有两个致命的性能瓶颈：产生的候选集过大，而且算法必须耗费大量的时间处理候选项集；多次扫描数据库，需要很大的I/O负载，时间和空间复杂度高。

为了提高算法的效率，Apriori算法运用了“频繁项集的子集是频繁项集，非频繁项集的超集是非频繁项集”这一性质有效地对频繁项集进行修剪。如果C_k中某个候选项集有一个(k–1)–子集不属于L_{k-1}，则这个项集可以被修剪掉，这个修剪过程可以降低计算所有候选集的支持度的代价[①]。

虽然Apriori算法已进行了一定的优化，但在实际的应用中仍然存在不足，于是人们相继提出了基于栈变换、基于划分、基于采样、基于Hash的算法等，来提高频集算法的效率。

6.2.3 FP Growth

针对Apriori算法的固有缺陷，FP Growth使用FP树的紧凑数据结构来组织数据，直接从该结构中提取频繁项集。

FP Growth的基本思想如下。

采取分而治之的策略，在保留项集关联信息的前提下，将数据库的频集压缩到一棵频繁模式树中；再将这种压缩后的FP树分成一些条件数据库并分别挖掘每个条件库。在算法中有两个关键步骤：一是生成频繁模式树FP–Tree；二是在频繁模式树FP–Tree上发掘频繁项集。

算法描述如下。

（1）对于每个频繁项，构造它的条件投影数据库和投影FP–Tree。

（2）对每个新构建的FP–Tree重复这个过程，直到构造的新FP–Tree为空，或者只包含一条路径。

（3）当构造的FP–Tree为空时，其前缀即为频繁模式；当只包含一条路径时，通过枚举所有可能组合并与此树的前缀连接即可得到频繁模式。

---

① http://www.4oa.com/article/html/6/32/475/2005/16702.html

6.2.4　多层关联规则

对于许多应用，由于多维数据空间数据的稀疏性，在低层或原始层的数据项之间很难找出强关联规则。现实生活中许多的概念都存在层次性，在进行数据挖掘时，可以引入相关概念层次，从而在较高的层次上进行挖掘。对不同的用户来说，信息的价值是不同的，所以，虽然在较高层次上挖掘得到的规则可能是更为普通的信息，但对于一些用户也许是非常有价值的信息。因此，数据挖掘应该具备在多个层次上进行挖掘的能力。多层关联规则可分为同层关联规则和层间关联规则。

同层关联规则可采用以下两种支持度策略。

（1）统一的最小支持度：对于不同的层次，使用相同最小支持度。此策略对于用户来说比较容易操作，而且算法也比较容易实现，但是也存在一定的弊端。

（2）递减的最小支持度：每个层次使用不同的最小支持度，较低层次使用的最小支持度相对较小，还可以利用在上层挖掘得到的信息进行一些相关的过滤工作。

层间关联规则应该根据较低层次的最小支持度来确定挖掘时使用的最小支持度。

6.2.5　多维关联规则

多维关联规则指涉及两个或两个以上层次的关联规则。根据同一个维在关联规则中是否重复出现，可以把多维关联规则分为两种类型：一种是维间关联规则，该规则只涉及相同的维，即维不重复出现；另一种是混合维关联规则，该规则涉及多个维，即维可以重复出现。比如“年龄20至30，喜欢郊游→喜欢游泳”就是混合维关联规则。这两种关联规则的挖掘还要考虑不同的字段种类，即类别型字段与数值型字段[①]。一般的数据挖掘算法都可以对类别型字段进行关联规则挖掘，而对数值型字段，就需要将其进行一定的处理才可以进行关联规则的挖掘。

处理数值型字段的方法有以下几种。

（1）数值型字段被分成一些由用户预定义的层次结构，然后对其进行关联规则挖掘，得出的规则叫做静态数量关联规则。

（2）根据数据的分布，数值型字段被动态地分成一些布尔字段。每个字段表示一个数值字段的区间，落在其中为1，反之为0。经挖掘得出的规则叫做布尔数量关联规则。

（3）考虑数据之间的距离因素，数值型字段被分成一些能体现它含义的区间。经挖掘得出的规则叫做基于距离的关联规则。

（4）直接分析数值型字段中的原始数据。运用相关的统计方法对其进行分析，同时结合多层关联规则的概念，在多个层次之间进行比较，得出的规则叫做多层数量关联规则。

6.3　分类与回归

数据库中隐藏着许多可以为商业、科研等活动的决策提供参考的知识。目前机器学习、模式识别、统计学和神经网络学等领域的研究人员提出了许多预测方法。其中分类和回归就

① http://www.4oa.com/article/html/6/32/475/2005/16702.html

是两种不同的预测方法，分类主要是用于预测离散的目标变量，输出的是离散值；而回归用于预测连续的目标变量，输出的是有序值或连续值。本节主要介绍分类与回归的概念，进行分类和回归时主要使用的技术和方法。

6.3.1　基本概念

分类任务就是确定对象属于哪个预定义的目标类。分类问题是一个普遍存在的问题，有许多不同的应用。例如，根据电子邮件的标题和内容检查出垃圾邮件，根据核磁共振扫描的结果区分肿瘤是恶性的还是良性的等。数据分类过程包括两个步骤，如图6.5所示。

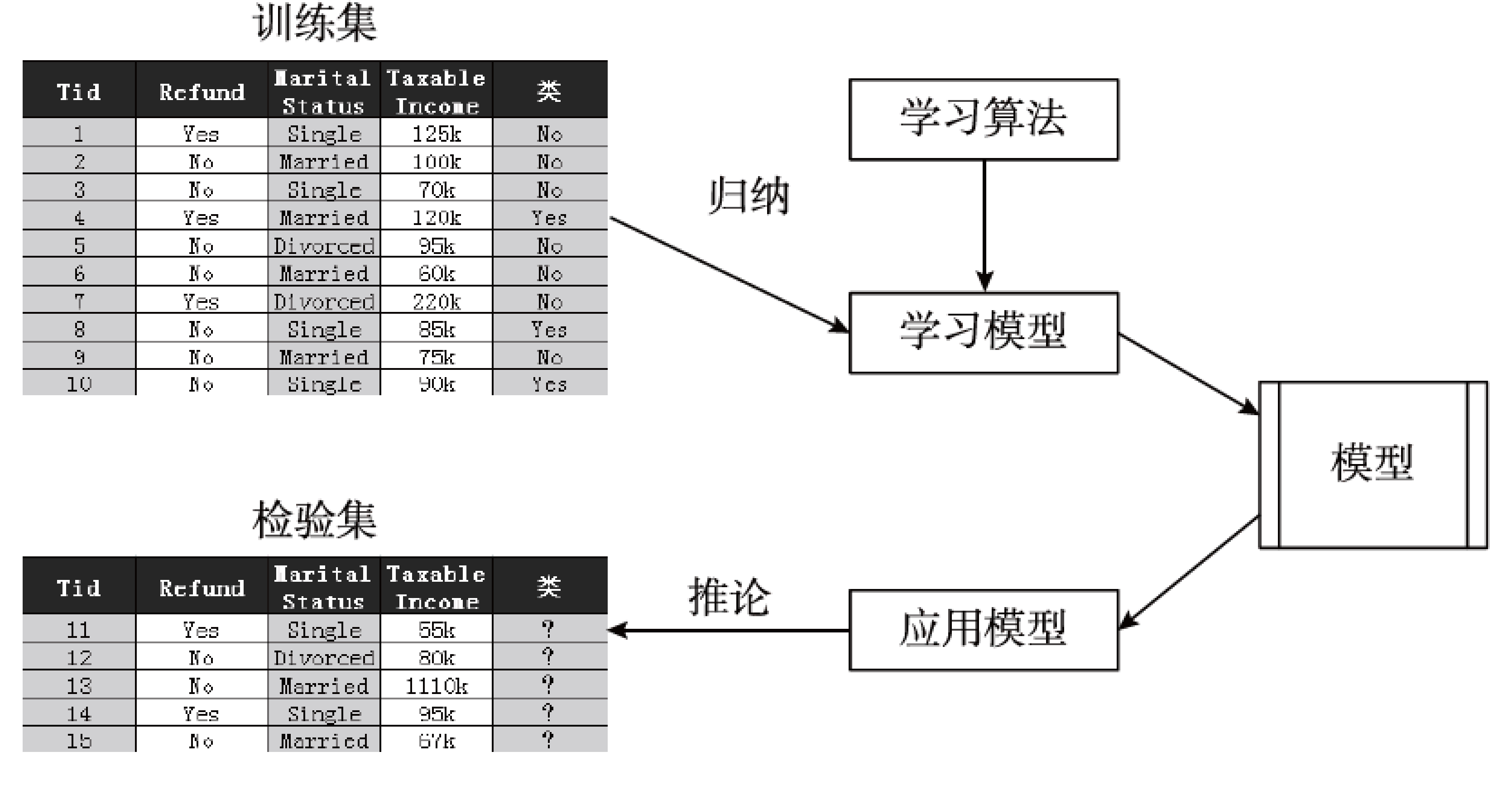


图6.5　建立分类模型的一般方法

第一步，建立一个已知数据集类别或概念的模型。通过分析属性所描述的数据库元组来构造模型。每一数据行都可认为是属于一个确定的数据类别，其类别值是由一个属性描述（被称为类标号属性）。为建立模型而被分析的数据集称为训练数据集，其中的单个元组称为训练样本，由样本群随机地选取。

分类学习又可称为监督学习，它是在已知训练样本类别情况下，通过训练建立相应模型；而无监督学习（聚类）则是在训练样本的类别与类别个数均未知的情况下进行的，那里每个训练样本的类标号都是未知的，要学习的类集合或数量也可能事先不知道。通常，分类学习模型以分类规则、判定树或数学公式的形式提供。

第二步，使用模型进行分类。先评估模型的预测准确率，当该准确率可以接受时，则可以用其对类标号未知的数据元组或对象进行分类。

分类模型的性能根据模型正确和错误预测的检验记录计数进行评估，这些计数存放在称作混淆矩阵的表格中。表6.2描述了一个二元分类问题的混淆矩阵。表中每个表项f_{ij}表示实际类标号为i被预测为j的记录数。

表6.2　二类问题的混淆矩阵

		预测的类	
		类=1	类=0
实际的类	类=1	f_{11}	f_{10}
	类=0	f_{01}	f_{00}

按照混淆矩阵中的表项，被分类正确预测的样本总数是（$f_{11}+f_{00}$），而被错误预测的样本总数是（$f_{10}+f_{01}$）。另外，使用准确率或其他性能度量来衡量分类模型的性能，定义如下：

$$准确率=\frac{正确预测数}{预测总数}=\frac{f_{11}+f_{oo}}{f_{11}+f_{10}+f_{01}+f_{00}} \tag{6-4}$$

$$错误率=\frac{错误预测数}{预测总数}=\frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}} \tag{6-5}$$

大多数分类算法都在寻找这样一些模型，当把它们应用于检验集时具有最高的准确率。

6.3.2　决策树

决策树是一个预测模型，它是一种由结点和有向边组成的树结构。其中，树的最顶层结点是根结点，每个内部结点表示属性的某个对象，每个分枝代表一个属性值输出，每个叶结点代表类或类分布。从根结点到叶结点的一条路径就表示一条合取规则，而整个决策树表示一组析取表达式规则。一棵典型的决策树如图6.6所示，该决策树描述了一个购买电脑的分类模型，其中一般内部结点用矩形表示，而树叶用椭圆表示。

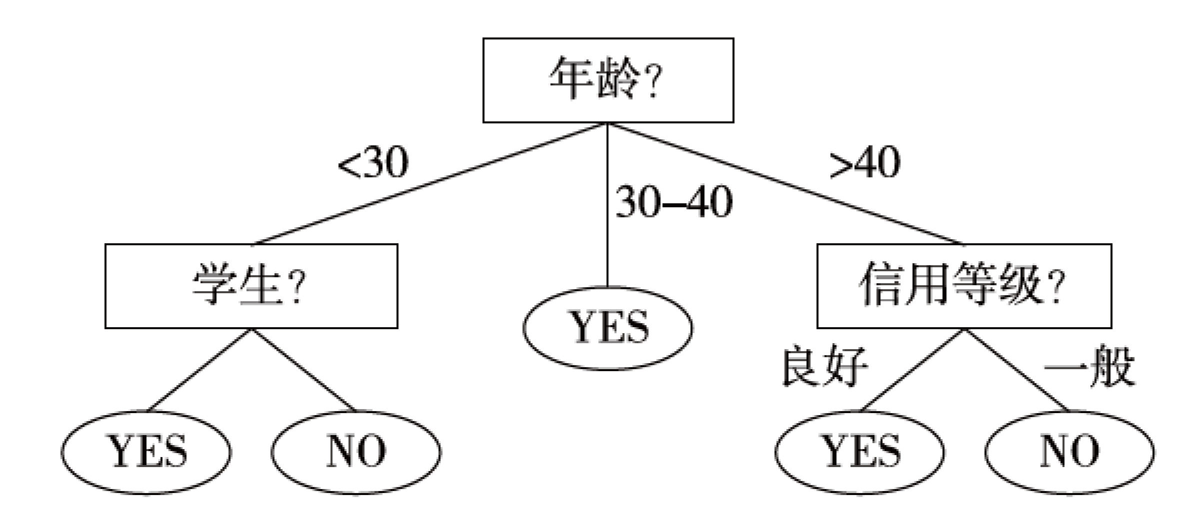


图6.6　决策树示意描述

著名的ID3算法是Quinlan在1986年提出来的，在此算法的基础上，1993年他又提出了C4.5算法，形成新的监督学习算法的性能比较标准。ID3算法和C4.5都是采用非回溯的方法。然而这两种方法已经越来越难以适应较大规模数据集的处理需求，因此在此基础上又提出了一些新的改进算法，其中SLIQ（Supervised Learning In Quest）和SPRINT（Scalable Parallelizable Induction of decision Trees）是两种比较有代表性的算法。

1. ID3算法

ID3算法的理论基础是信息论，它的核心思想是：用信息增益（information gain）作为属性选择的衡量标准，在决策树各级结点上选择属性，使得在每一个非叶结点进行测试时，能获得关于被测试记录最大的类别信息。其具体操作方法是：检测所有的属性，选择信息增益最大的属性产生决策树结点，由该属性的不同取值建立分支，再对各分支的子集递归调用该方法建立决策树结点的分支，直到所有子集仅包含同一类别的数据为止。最后得到一棵决策树，可以使用它对未知的类别、新的样本进行分类。

信息增益的计算方法是：按照上述方法计算每个属性的信息增益，并比较它们的大小，这样就能很容易地获得具有最大信息增益的属性。

设S是s个样本的集合，假定类标号属性具有m个不同值，定义m个不同类$C_i(i=1,2,\cdots,m)$。设s_i是C_i中的样本数，对一个给定的样本分类所需的期望信息由下式给出：

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (6\text{-}6)$$

其中，p_i表示任意样本属于C_i的概率，并用si/s估计。注意，对数函数以2为底，因为信息用二进位编码。

设属性A具有v个不同值$\{a_1, ..., a_v\}$。可以用属性A将S划分为v个子集$\{S_1, ..., S_v\}$；其中，S_j包含S中这样一些样本，它们在A上具有值a_j。如果A选作测试属性（即最好的划分属性），则这些子集对应于由包含集合S的结点生长出来的分枝。设s_{ij}是子集S_j中类C_i的样本数。根据A划分子集的熵或期望信息由下式给出：

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}) \quad (6\text{-}7)$$

项$\frac{s_{1j} + \cdots + s_{mj}}{s}$充当第$j$个子集的权，并且等于子集（即，$A$值为$a_j$）中的样本个数除以$S$中的样本总数。熵值越小，子集划分的纯度越高。对于给定的子集S_j，其信息期望为：

$$I(s_{1j}, s_{2j}, \cdots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \quad (6\text{-}8)$$

其中，p_{ij}表示S_j中的样本属于C_i的概率，用s_{ij}/s_j表示。

在属性A上分枝将获得的信息增益是：

$$\text{Gain}(A) = I(s_1, s_2, ..., s_m) - E(A) \quad (6\text{-}9)$$

ID3算法的优点是：算法的理论清晰，方法简单易操作，学习能力较强。其缺点是：对比较大的数据集失效，对噪声也比较敏感；当训练数据集加大时，产生的决策树可能会随之改变。

2. C4.5算法

C4.5算法是在ID3算法的基础上改进而来的，具体的改进包括以下几个方面。

- 选择属性时，用信息增益率代替信息增益，克服了用信息增益选择属性时偏向选择取值较多的属性的缺点。
- 在构造树的过程中直接进行剪枝。
- 实现了连续属性的离散化处理。
- 实现了不完整数据的处理。

与其他分类算法，如统计方法、神经网络等比较，C4.5算法的优势是产生的分类规则易于理解，准确率较高。但是它也存在一些缺点，如在树的构造过程中，需要对数据集进行多次的顺序扫描和排序，算法效率比较低；此外，当训练集大到无法在内存容纳时，C4.5算法无法运行，此算法只适用于可以驻留于内存的数据集。

3. SLIQ算法

SLIQ算法是IBM于1996年提出的，在C4.5算法的基础上进行了改进，它是一种高速可调节的数据挖掘分类算法，主要解决当训练集数据量巨大，无法全部放入内存时，进行高速、准确的生成树的问题。它在构造决策树的过程中采用了“预排序”和“广度优先策略”两种技术。

- 预排序。在以前的算法中对于连续属性在每个内部结点寻找其最优分裂标准时，都要对训练集按照该属性的取值进行排序，而排序是很费时的操作。为了降低数值型属性的排序代价，提高处理速度，SLIQ算法采用了预排序技术。预排序技术就是针对每

个属性的取值，把所有的记录按照从小到大的顺序进行排序，从而避免了在决策树的每个结点对数据集进行的排序操作。在操作时，需要为训练数据集的每个属性创建一个属性列表，为类别属性创建一个类别列表。属性表可以写回磁盘。

- 广度优先策略。C4.5算法中的树的构造是采用深度优先策略完成的，此策略需要对每个属性列表在每个结点处都进行一遍扫描以完成结点的分裂，会消耗大量的时间。为了节省操作时间，提高运行效率，SLIQ算法利用广度优先策略生成决策树，此策略对每层结点只需要扫描一次属性列表，就可以找到决策树中每个叶子结点的最优分裂方式。

与C4.5算法相比，SLIQ算法采用了上述两种技术，能够处理比C4.5规模大得多的训练集，且随着记录个数和属性个数的增长，在一定范围内具有较好的可伸缩性。

但是，SLIQ算法也存在以下缺点：

- 该算法创建的类别列表需要存放于内存，而类别列表的元组数与训练集的元组数是相同的，这在一定程度上限制了能够处理的数据集的大小。
- 该算法采用了预排序技术，而排序算法的复杂度并不与记录数目成线性关系，这使得该算法不可能达到随记录数目增长的线性可伸缩性。

4. SPRINT算法

为了减少滞留在内存中的数据量，SPRINT算法对决策树算法的数据结构进行了进一步的改进，即删除了在SLIQ算法中需要存放于内存的类别列表，将它的类别列合并到每个属性列表中，这样就不必参考其他信息。在遍历每个属性列表寻找当前结点的最优的分裂标准时，对结点的分裂就表现在对属性列表的分裂，即将每个属性列表分裂成两个列表，分别存放各个结点的记录。

SPRINT算法可以更为简便地寻找到每个结点的最优分裂标准，但其也存在缺点，即分裂非分裂属性的属性列表则难以实现。要克服这一缺点，可以在对分裂属性进行分裂时用哈希表记录每个记录属于哪个子结点，如果内存可以容纳整个哈希表，则其他属性列表的分裂只需要参考该哈希表。哈希表的大小与训练的数据集的大小成正比，当训练的数据集很大时，内存也许不能容纳哈希表，而要分批执行分裂操作，这使得SPRINT算法不具有很好的可伸缩性。

6.3.3　贝叶斯分类算法

贝叶斯分类算法是统计学的一种分类方法，是一类利用概率论、统计学等知识进行分类的算法，如朴素贝叶斯（Naive Bayes，NB）算法、树增强型朴素贝叶斯（TAN）算法等。对于一个未知类别的样本，贝叶斯分类算法首先运用贝叶斯定理来预测该未知类别的样本属于各个类别的可能性，然后比较可能性的大小，将可能性最大的一个类别确定为该样本的最终类别。因为贝叶斯定理是在一个很强的独立性假设前提下才成立的，而此假设在大多数实际情况下是不成立的，所以这在一定程度上降低了此算法的分类准确性。为了克服这一缺点，增强分类的准确性，很多降低独立性假设的贝叶斯分类算法应运而生，TAN（Tree Augmented Bayes Network）算法就是其中的一种。

1. NB算法

设每个数据样本用一个n维特征向量来描述n个属性的值，即$X=\{X_1,X_2,\cdots,X_n\}$，假定有m个类，分别用$C_1,C_2,\cdots,C_m$表示。给定一个未知的数据样本X，若NB算法将未知的样本X分配给类C_i，则必有：

$$P(C_i|X)>P(C_j|X);1\leqslant j\leqslant m,i\neq j \quad (6\text{–}10)$$

根据贝叶斯定理：

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)} \quad (6\text{–}11)$$

如果训练数据集有较多属性和元组，则$P(X|C_i)$的计算量可能非常大。为减少计算量，通常情况下都假设各属性的取值是互相独立的，则：

$$P(X|C_i)=\prod_{k=1}^{n}P(x_k|C_i) \quad (6\text{–}12)$$

可以从训练数据集计算得到先验概率$P(x_k|C_i)$。对一个未知类别的样本X，运用NB算法，先分别计算出该样本X属于每一个类别C_i的概率$P(X|C_i)P(C_i)$，然后比较概率值的大小，将概率最大的类别作为该样本的类别。

当满足各属性之间相互独立的前提时，NB算法才成立。当数据集满足这种较强的独立性假设时，该算法分类的准确性较高，否则准确性可能较低。除此之外，该算法不输出分类规则。

2. TAN算法

TAN算法在NB网络结构的基础上增加属性对之间的关联（边），通过发现属性对之间的依赖关系，从而有效地降低了NB算法中各个属性之间相互独立的假设。

该方法实现的步骤是：用结点表示属性，有向边表示各个属性之间的依赖关系；把类别属性作为根结点，其余所有的属性作为根结点的子结点。一般情况下，用虚线表示NB所需要的边，用实线表示新增加的边。属性A_i与A_j之间的有向边表示它们之间存在着依赖关系，即属性A_i对类别变量C的影响还取决于属性A_j的取值。

增加的边需要满足以下条件：类别变量没有双亲结点，每个属性有一个类别变量双亲结点和最多另外一个属性作为其双亲结点。

确定这组关联边后，就可以计算一组随机变量的联合概率分布如下：

$$P(A_1,A_2,\cdots,A_n,C)=P(C)\prod_{i=1}^{n}P(A_i|\prod A_i) \quad (6\text{–}13)$$

其中$\prod A_i$表示A_i的双亲结点。TAN算法考虑了n个属性中$(n-1)$个两两属性之间的关联性，在一定程度上降低了对属性之间独立性的假设，但由于没有考虑到属性之间可能存在的其他关联性，因此，在很大程度上限制了该算法的适用范围。

3. 贝叶斯网络

贝叶斯网络是用来表示变量间连接概率的图形模式，它表示多个指标的联合分布，提供了一种自然的表示因果信息的方法，用来发现数据间的潜在关系。贝叶斯网络是一种有向无环图（Directed Acyclic Graph，DAG），图中结点表示随机变量，结点间的有向边表示变量间的概率依赖关系，依赖的强度或者说不确定性通过条件概率来表示。

贝叶斯网络可以表示成在随机变量集$X=\{X_1,X_2,\cdots,X_n\}$上的一个二元组，用符号G(B,P)表示，变量X_i的取值可记为x_i，变量间的条件独立性可表示为：$I(X_i; X_j \mid X_K)$，其含义为在给定的X_K的条件下，X_i与X_j相互独立。其中G(B,P)由两部分构成：

（1）G代表一个具有n个结点的有向无环图，结点对应随机变量$X=\{X_1,X_2,\cdots,X_n\}$，结点变量可以是任何问题的抽象用以代表属性、状态、测试值等。结点之间的有向边（弧）反映了变量间的依赖关系，有向弧箭头指向的结点称为子结点，箭尾对应的结点称为父结点。

（2）P表示给定父结点条件下，每个结点的条件概率分布（Conditional Probability Distributing，CPD）条件概率分布，可以用$P(X_i \mid Pa(X_i))$来描述，其中$Pa(X_i)$表示结点X_i的父节点集合。

依据马尔科夫独立性假设，贝叶斯网络规定图中的任一结点X_i条件独立于由X_i的父结点给定的非X_i后代结点构成的任何结点子集，形式上可记为：$\forall i$，$I(X_i; NonDescedents(X_i) \mid Pa(X_i))$，式中$NonDescedents(X_i)$表示$X_i$的非子节点。有向无环图$G$即代表了一系列的条件独立假设，这些假设使得随机变量$X=\{X_1,X_2,\cdots,X_n\}$的联合概率分布具有可分解性，这大大减少了确定联合分布的参数的数目，联合分布的分解式可如下表示：

$$Pr\{X_1,X_2,\cdots,X_n\}=\prod_{i=1}^{n}Pr\{X_i|Pa(X_i)\} \qquad (6\text{-}14)$$

通过引入独立关系，联合概率分布可以被分解成更小的因式，从而简化知识获取和先验知识的建模过程，降低计算复杂度。

贝叶斯网络的结构G表示了在马尔可夫独立性假设条件下的一组独立性假设。令Ind(G)表示一组条件独立关系（如$I(X_i; X_j \mid X_K)$），在贝叶斯网络中，往往有不同的有向无环图可以表示相同的条件独立关系，即Ind(G)=Ind/(G')，这些图称为贝叶斯网络等价类。贝叶斯网络等价类可用部分有向图（partially directed graph，PDAG）表示，PDAG中有向边$X\rightarrow Y$表示贝叶斯网络等价类中所有的DAG均含有此有向边，PDAG中的无向边$X\rightarrow Y$表示，贝叶斯网络等价类中部分DAG的边的方向为$X\rightarrow Y$，部分为$X\leftarrow Y$，所有的PDAG称为CPDAGs(Completed–PDAGs)，它包含了贝叶斯网络中的所有等价类。

6.3.4　人工神经网络

人工神经网络（Artificial Neural Network，ANN）是人们在模拟人脑处理问题的过程中发展起来的一种新型的信息处理理论，也称神经网络，是由大量类似神经元相互连接构成的非线性的复杂系统。神经网络基于模仿生物大脑的结构和功能，采用数学和物理方法进行研究而构成的一种信息处理系统，通过调整各神经元之间的连接强度，模拟人的学习、归纳和分类能力，广泛应用于信号处理、模式识别、智能检测及其他工程领域。

1. 人工神经元模型

常用的人工神经元模型主要基于模拟生物神经元信息的传递特性，即输入、输出关系。如果用模拟电压表示生物神经元输入、输出脉冲的密度，则生物神经元信息传递的主要特性可以用图6.7的模型来模拟。

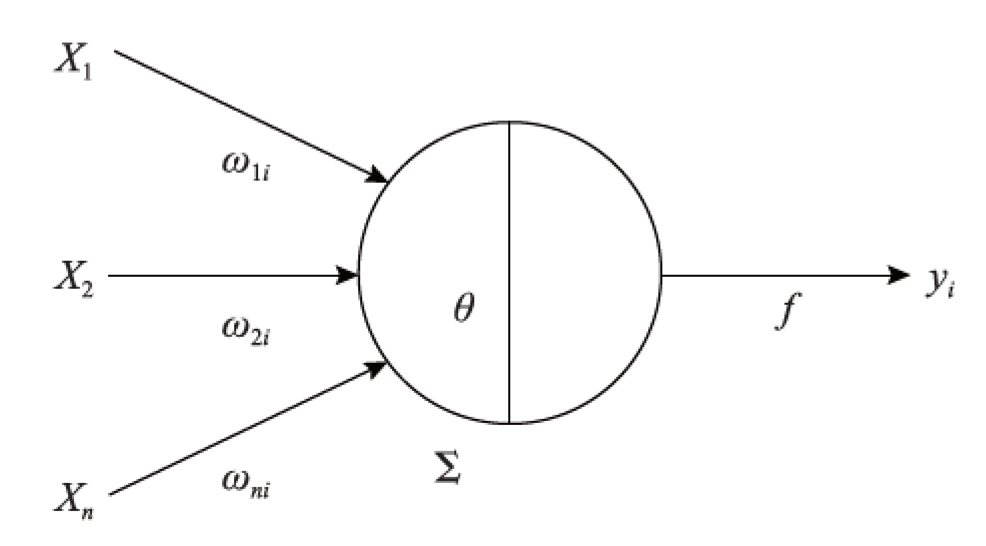


图6.7　人工神经元模型

在图6.7中，X_1（i=1,2,⋯,n)表示加于输入端突触上的输入信号；ω_i表示相应原突触连接权重系数，它是模拟突触传递强度的一个比例系数；Σ表示突触后信号的空间累加；θ表示神经元的阈值；f表示神经元的响应函数。

此人工神经元模型的数学表达式为：

$$I_i=\sum_{j=1}^{n}\omega_{ji}x_j-\theta_i \tag{6-15}$$

$$y=f(I_i) \tag{6-16}$$

传递函数$f(x)$通常是非线性函数，比如阶跃函数、S状曲线等，也可以是线性函数。以下是一些常用的神经元非线性函数。

（1）阈值型函数

此模型中，神经元没有内部状态，当y_i取0或1时，$f(x)$为一阶跃函数：

$$f(x)=\begin{cases}1, x\geqslant 0\\0, x<0\end{cases} \tag{6-17}$$

当y_i取−1或1时，$f(x)$为sgn函数：

$$\mathrm{sgn}(x)=\begin{cases}1, x\geqslant 0\\-1, x<0\end{cases} \tag{6-18}$$

（2）S型函数

通常为在（0，1）或（−1，1）内连续取值的单调可微分的函数。常用响应函数为S型（Sigmoid）函数：

$$f(x)=\frac{1}{1+e^{-\beta x}},(\beta>0) \tag{6-19}$$

当β趋于无穷时，S状曲线趋近于阶跃函数。一般情况下，β取值为1。

2. 人工神经网络的构成

当神经元的模型确定之后，一个神经网络的特性和能力主要是由网络的拓扑结构及学习方法决定的。按照不同的神经元互联模式，神经网络主要可分为前馈网络、反馈网络、自组织网络和随机型网络四类。

- 前馈网络：网络可划分为若干层（包括输入层、隐含层和输出层），一般情况下，第i层的神经元只接收第（i−1）层神经元传递的信号，各神经元间没有反馈。
- 反馈网络：反馈网络在前馈网络的基础上增加了反馈连接，可以是异反馈也可以是自反馈，同时网络中还可以有计算功能的隐神经元。

- 自组织网络：该网络能够识别环境中的特征，并自动聚类。
- 随机型网络：在网络运行和学习算法中引入的随机机制。

3. 人工神经网络的学习

当神经网络的拓扑结构确定后，需要有相应的学习方法与它配合，它才会具备某些智能特性。因此，人工神经网络研究的核心问题是学习方法。

对于人工神经网络，它的适应性是通过学习实现的，学习是神经网络研究的一个重要内容。学习方法实质上就是网络连接权的调整方法。人工神经网络一般通过以下两种方法确定连接权值：一种是根据具体要求直接计算出来的，比如Hopfield网络的优化计算就是运用这种方法；另一种是通过不断学习调整得到的，大多数人工神经网络都运用这种方法。

由于网络结构和功能的不同，学习方法也是各种各样的，这里仅仅介绍人工神经网络中一些比较基本的、常用的学习规则。

- 修正型。这属于一种有导师的学习规则，连接权调整的依据是理想处理的结果与实际处理结果的误差值。通过不断地减小误差来完成整个学习过程。
- Hebb型。这条规则的思路是通过连接神经元的状态来改变权值。神经元只有两种状态："兴奋"和"抑制"。当相邻的两个处理单元同时处于"抑制"状态时，连接变弱。
- 随机型。此条规则中，网络变量的调整是以激励函数的改变为准的。其中的变化带有一定的随机性。此时网络变量可以是权值，也可以是处理单元的状态。
- 竞争型。是指在无导师的学习过程当中，只提供训练的样本，而不设计输出的结果，处理单元通过自身的学习，进入到相应的类别当中。处理单元通过相互的竞争，只留下几个处理单元处于"兴奋"的状态，以感应输入信号的影响。

神经网络的主要任务是对外部世界进行建模，并通过学习使模型与外部环境充分一致，来完成特定的任务。不同的网络类型使用不同的学习方法可以获得不同的网络模型，在实际应用中，需要根据具体应用的特点来设计合适的网络模型。

在网络学习阶段，为了实现输入样本与其相应正确类别的对应，网络需要调整权重值。神经网络的学习之所以也叫做连接学习，是因为网络主要是针对其中的连接权重进行学习的。图6.8表示的是一个典型的多层前馈神经网络。

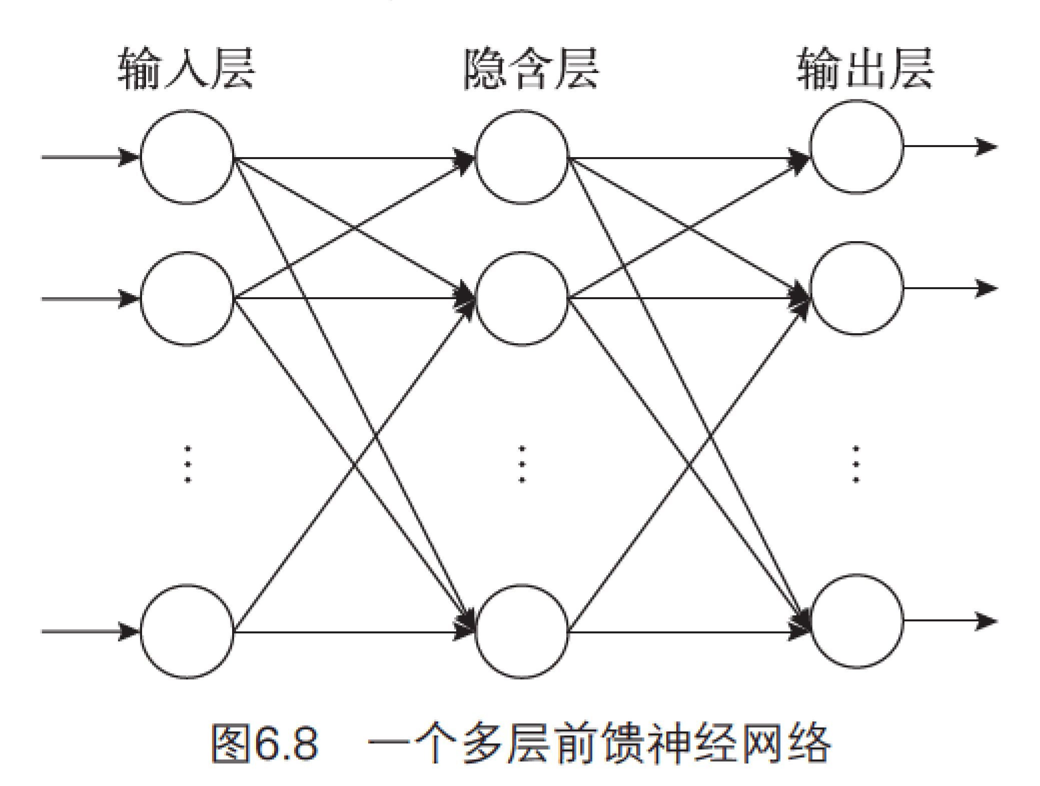


图6.8　一个多层前馈神经网络

由于人工神经网络学习时间比较长，因此它仅仅适用于时间容许的应用场合，这在一定

程度上限制了它的应用。此外它还需要一些关键参数，如网络结构等，这些参数通常需要经验才能够有效地确定；且神经网络的输出结果也是比较难理解的。

人工神经网络的优点就是对未知的数据具有较好的预测、分类能力，而且对噪声数据有较好的适应能力。目前不断有人提出一些从神经网络中挖掘出（知识）规则的新算法。

6.3.5 支持向量机

支持向量机（Support Vector Machine，SVM）方法是建立在统计学习理论的VC维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以求获得最好的推广能力。

支持向量机是在线性可分条件下的最优分类面发展起来的，对于常见的二分类问题，假设有训练集（x_i, y_i）,i=1,2,⋯,n,$x_i \in R^d$,$y \in \{-1,1\}$，可以被一个超平面分开，如图6.9所示，两种图标代表两种类别，h：$(w \times x_i)+b=0$表示分类线，两类的边界样本都在h_1，h_2两条直线上，且两条直线都是与h平行的。两条直线的之间的距离表示两类样本点之间的距离。最优h既能将两个类别毫无偏差的区分，又能使分类间隔最大。

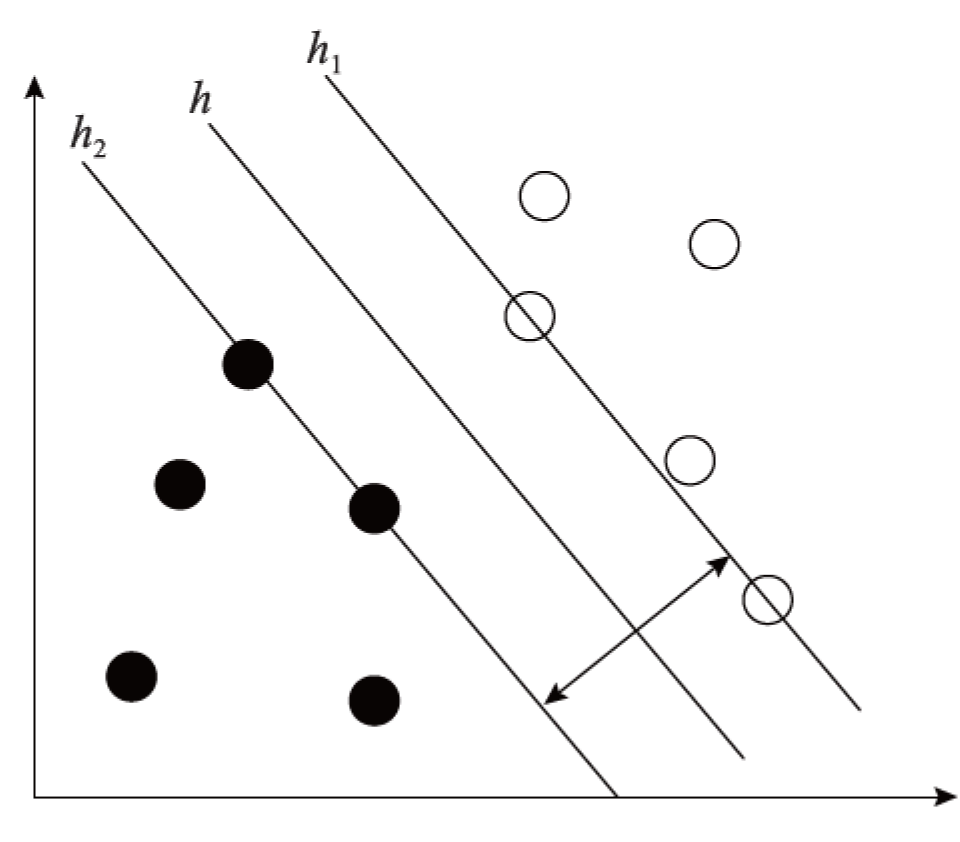


图6.9　最优分类超平面

可以用以下公式表示直线h_1，h_2：

$$y_i[(w \times x_i)+b]-1=0, i=1,2,...,n \tag{6-20}$$

两类样本之间的分类距离margin$=2/\|w\|$，其中使$\|w\|^2$最小的分类线就是最优分类线。支持向量是指两类样本中离分类面最近的超平面上的点，即h_1，h_2上的点就是支持向量。

1. 线性可分SVM

由上可知，超平面$(w \times x)+b=0$能将两类样本正确区分开来，使得分类间隔最大的优化问题就表示为：

$$\min \varphi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}w^T w$$

满足

$$y_i[(w \times x_i)+b]-1 \geqslant 0, i=1,2,...,n \tag{6-21}$$

公式中约束条件和目标函数都满足凸条件，故其具有惟一的全局解且值最小。为了导出

原始问题的对偶问题，引入了Lagrange函数：

$$L(w, b, a)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}a_i\{y_i[(w\times x_i)+b]-1\} \quad (6\text{–}22)$$

其中$a=(a_1,a_2,\ldots,a_n)^T$是Lagrange乘子向量。因此，我们得到原问题的对偶问题：

$$\min\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_i y_j(x_i\times x_j)a_i a_j-\sum_{j=1}^{n}a_j$$

满足

$$\sum_{i=1}^{n}a_i y_i=0,\ a_i\geqslant 0,\ i=1,2,\cdots,n \quad (6\text{–}23)$$

求解上面的凸二次规划问题可以得到解$a^*=(a_1^*,\cdots,a_n^*)^T$，计算$w^*=\sum_{i=1}^{n}a_i^* y_i x_i$，选取$a^*$的一个正分量$a_j^*$，据此计算：

$$b^*=y_i-\sum_{i=1}^{n}a_i^* y_i(x_i\times x_j) \quad (6\text{–}24)$$

构造分化超平面（$w^*\times x$）$+b^*=0$，由此求得决策函数，

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}y_i a_i^*(x_i\times x)+b^*) \quad (6\text{–}25)$$

2. 线性不可分SVM

有些样本集在线性条件下是不能够被正确分类的。但是可以放宽正确分类的条件，只要这个样本点落在能够被正确分类点的附近，就可以认为这个样本点能够被“正确”分类。这就是引入松弛变量的思想，那么问题的表述形式变成：

$$y_i[(w\times x_i)+b]-1+\xi_i\geqslant 0;\ \xi_i\geqslant 0,\ i=1,2,\cdots,n \quad (6\text{–}26)$$

寻找目标函数变为：

$$\varphi(w)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i \quad (6\text{–}27)$$

为了避免ζ_i取值太大降低解的质量，引入了可调参数C对没有正确分类样本进行惩罚，C越大对错误的惩罚越重。常用Lagrange方法将上式最优分类面问题转化为对偶问题，在$\sum_{i=1}^{n}y_i a_i=0$和$0\leqslant a_i\leqslant C, i=1, 2,\cdots,n$的约束下，用与线性可分情况下相同方法求解这一优化问题，同样得到一个二次函数极值问题，得到最优分类判别函数：

$$f(x)=\mathrm{sgn}\{w^*\times x+b^*\}=\mathrm{sgn}(\sum_{i=1}^{n}y_i a_i^*(x_i\times x)+b^*) \quad (6\text{–}28)$$

3. 非线性SVM

非线性SVM的基本思想是通过事先确定的非线性映射将输入向量x映射到一个高维特征空间（Hilbert空间）中，然后在此高维空间中构建最优超平面。

从上一节对线性SVM的讨论中可以看出，向量之间只进行点积运算。那么映射到的高维空间中也只需要内积运算，但是在高维空间进行内积运算是一个很繁琐的过程。如果能够在低维空间就能计算出高维内积，那么可以减少很多工作。研究人员发现存在某种类型的函数能够刚好满足上面的条件。

根据Hilbert–Schmidt原理，只要一种核函数满足Mercer条件，即对任意对称函数$K(x, x')$，它是某个特征空间中的内积运算的充要条件为：对任意的$\varphi(x)\neq 0$且，有：$\int\varphi^2(x)dx<\infty$

$$\iint K(x, x')\varphi(x)\varphi(x')dxdx'>0 \tag{6-29}$$

此时，二次规划问题的目标函数变为：

$$Q(a)=\sum_{i=1}^{n} a_i-\frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j(x_i\times x_j) \tag{6-30}$$

对应的分类决策函数可以表示为：

$$f(x)=\mathrm{sgn}\{w\times x+b\}=\mathrm{sgn}(\sum_{i=1}^{n} y_i a_i^*(x_i\times x)+b^*) \tag{6-31}$$

4. 核函数

支持向量机的关键在于核函数。低维空间向量集通常难于划分，解决的方法是将它们映射到高维空间。但是这个办法带来的困难就是计算复杂度的增加，而核函数正好巧妙地解决了这个问题。也就是说，只要选用适当的核函数，就可以得到高维空间的分类函数。在支持向量机理论中，采用不同的核函数将导致不同的算法。

选择不同形式的核函数就可以生成不同的支持向量机，常用的有以下几种。

（1）线性核函数：$K(x,y)=x\times y$。

（2）多项式核函数：$K(x,y)=[(x\times y)+1]^d$，d为参数，用该核函数的SVM是一个多项式分类器。

（3）高斯核函数：$K(x,y)=\exp(-\|x-y\|^2/(2\delta^2))$，用该核函数的SVM是一个径向基函数分类器。

（4）Sigmoid核函数：$K(x,y)=\tanh(k(x\times y)-\delta)$，用该核函数的SVM是一个两层的多层感知器神经网络。

由于确定核函数的已知数据也存在一定的误差，考虑到推广性问题，因此引入了松弛系数以及惩罚系数两个参变量来加以校正。在确定了核函数基础上，再经过大量对比实验等将这两个系数取定，该项研究就基本完成，适合相关学科或业务内应用，且有一定能力的推广性。

当然误差是绝对的，不同学科、不同专业的要求不一。目前，支持向量机已经在许多领域（生物信息学，文本分类和手写识别等）都取得了成功的应用。

6.3.6　其他分类方法

除了上述分类方法以外，常见的其他分类方法还包括k–最临近分类、基于案例的推理、遗传算法、粗糙集、模糊集等，下面简单介绍这五种分类方法。

1. k–最临近分类

最临近分类是一种基于类比学习的分类方法。设用n维数值属性描述训练样本，每个样本对应n维空间的一个点，则所有的训练样本都包含在n维空间中。对于一个给定的未知样本，k–最临近分类法首先在n维模式空间中进行搜索，然后按照接近程度选择k个训练样本，这k个训练样本是最接近给定的未知样本的，即为未知样本的k个“近邻”。通常用欧几里德距离定义“临近性”。设空间中两个点$X=(x_1,x_2\cdots,x_n)$和$Y=(y_1,y_2,\cdots,y_n)$，则它们的欧几里德距离为：

$$d(X, Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{6-32}$$

在这k个最接近未知样本的训练样本中，选择最公共的类作为未知样本的类。当$k = 1$时，则选择与未知样本最临近的训练样本的类作为未知样本的类。

当存放的训练样本数量很大时，该方法的计算开销较大。最临近分类可以返回给定的未知样本实数值预测，这是最临近分类的预测作用。在这种情况下，最临近分类返回的是与未知样本的k个最临近训练样本实数值标号的平均值。

2. 基于案例的推理

基于案例的推理（CBR）分类法是一种基于要求的分类方法。该方法存放的训练样本是较为复杂的符号描述，这与最临近分类法将训练样本作为欧氏空间中的点存放的方式不同。CBR的应用比较广泛，在商务、工程和法律领域等都有应用。CBR的商务应用包括顾客服务台问题求解、案例描述产品有关的诊断问题等。CBR在工程和法律方面的案例分别是技术设计和合法规则。

对于一个给定的待分类的未知案例，CBR首先检查是否有一个相同训练样本存在，若找到，则返回训练样本中所包含的解决方法；若没有相同训练样本存在，就寻找与待分类的未知例的组成有相似之处的训练样本，从某种意义上讲，这些训练样本也是新示例的最近邻。如果案例可以用图来表示的话，那么这就涉及到与未知案例相似的子图的搜索。基于案例的推理试图对最近邻的训练案例进行合并以给出一个（针对新案例）解决方法。若各案例返回方法不兼容，必要时还必须回溯搜索其他的解决方法。基于案例的推理器可以利用背景知识和问题求解策略来帮助获得一个可行的解决方法。

基于案例的推理分类方法中，存在的问题包括：寻找相似性度量方法（如子图匹配）、开发快速索引技术和求解方法的合并等。

3. 遗传算法

遗传算法是一种借鉴自然进化基本思想的随机化搜索方法。它是借鉴了进化生物学中的一些现象（包括遗传、突变、自然选择、杂交等）而发展起来的。该算法具有以下特点：直接作用于结构对象，对函数的连续性以及函数的求导没有限制；存在内在的并行性；具备较好的全局寻找最优解的特性等。遗传算法运用概率化的寻求最优解的方法，不必制定明确的规则，能在搜索过程中自动获取和积累有关搜索空间的知识，并自适应地控制搜索过程以求得最优解。同时对搜索方向进行自适应地调整。

遗传算法一般的操作步骤如下[①]。

（1）首先进行初始化操作。初始进化代数计数器设为t=0，最大进化代数设为T，初始群体记为$P(0)$，该初始群体是随机生成的，其中包含有M个个体。

（2）个体评价操作。通常用适应度对个体进行评价，该步骤就是对群体$P(t)$中的每个个体的适应度进行计算。

（3）选择运算操作。该操作将选择作用在群体上的算子。选择的目的是把优化的个体遗传到下一代或者通过配对交叉产生新的个体，然后再遗传到下一代。该操作是建立在个体评价操作上的，即是在个体适应度评估的基础上进行的。

① http://baike.baidu.com/view/45853.htm?func=retitle

（4）交叉运算操作。将交叉算子作用于群体，从而生成新的个体。交叉算子在该操作中具有关键性的作用，也是遗传算法的核心。

（5）变异运算操作。改动一些群体中个体串的某些基因座上的基因值，也就是在群体中运用变异算子。

下一代群体P(t+1)就是在群体P(t)经过选择、交叉、变异操作后得到的。

（6）算法终止条件的判断。算法在运行的过程中，会对终止条件进行判断，若满足终止条件，遗传算法操作终止。即若t>T，则算法终止，并输出最优解。最优解就是在进化过程中得到的具有最大适应度的个体。

遗传算法是计算机科学人工智能领域中用于解决最优化的一种启发式搜索算法，是进化算法的一种，这种启发式通常用来生成有用的解决方案来优化和搜索问题。遗传算法在适应度函数选择不当的情况下有可能收敛于局部最优，而不能达到全局最优。

4. 粗糙集

粗糙集理论可以应用于分类问题，以帮助找出噪声数据或不准确数据中所存在的结构关系。它只能处理离散值属性，而连续值属性必须在进行离散化后才能运用粗糙集理论进行处理。

对于给定的训练数据集，首先建立数据内部的等价类。形成等价类的所有数据样本是不作任何区分的，也就是说，对于描述数据的属性，这些样本是等价的。粗糙集理论就是基于这些等价类的建立。通常情况下，在给定的数据中，有些类不能精确地被可用的属性区分。在这种情况下，可以运用粗糙集理论对这些类进行近似地定义。给定类C的粗糙集定义用两个集合近似：C的下近似和C的上近似。根据关于属性的知识，如果数据样本明确属于C，则这些数据样本组成C的下近似；如果数据样本不可能被认为不属于C，则所有这些数据样本组成C的上近似，如图6.10所示，其中一个矩形区域表示一个等价类。判定规则可以对每个类产生，这些规则一般用判定表来表示。

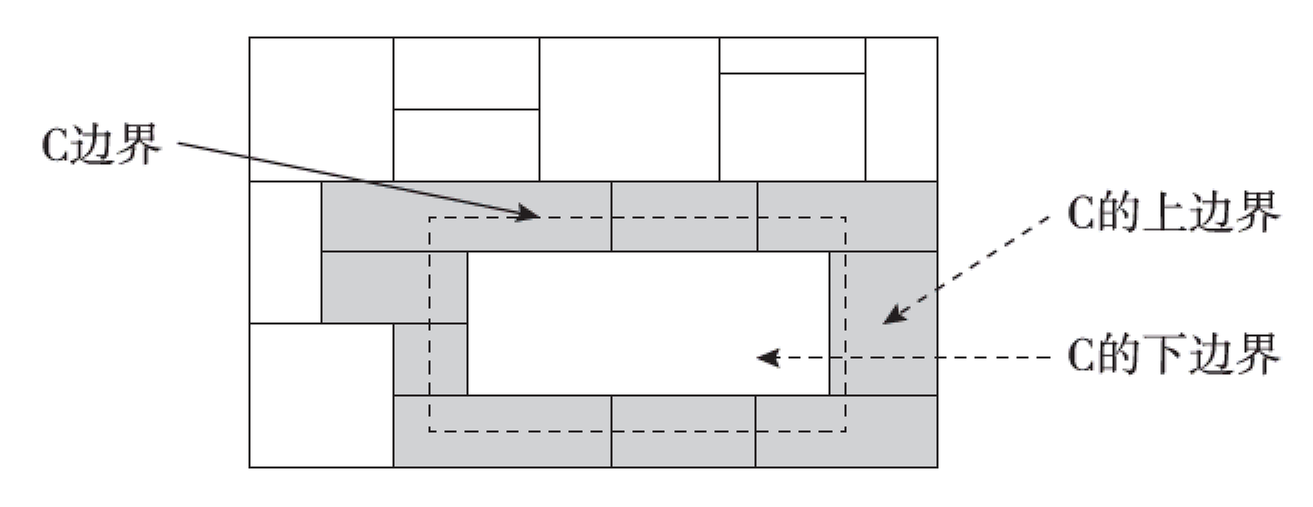


图6.10　一个粗糙集的示意图

可以运用粗糙集来进行特征归约、相关分析等操作，从给定的数据集中寻找出能够描述所有数据特征概念的最小属性集合本身就是一个NP–难问题。人们提出了一些可以帮助减少其计算复杂度的算法，例如利用可分辨矩阵，该矩阵储存每对数据样本之间属性取值的差别信息。借助可分辨矩阵就无需搜索这个数据样本集合，而只需要搜索该矩阵，就可以帮助发现冗余属性。

5. 模糊集

基于规则的系统应用于分类，处理连续值时是间断的，这是基于规则分类存在的不足。

例如对于顾客信用申请批准，应用以下规则：批准一个工作时间为二年或二年以上且收入较高（如：*income*50≥*K*）的人的信用申请。

$$IF(years_employed \geq 2) \wedge (income \geq 50K) THEN \cdots credit = approved$$

由以上规则，一个工作时间为二年以上的顾客，若他的收入大于50K，则他的信用申请将被批准，但若他的收入为49K，则他的申请就得不到批准。

这样的处理显然是不合理的。在这种情况下，引入模糊逻辑有利于解决这一问题。由于模糊逻辑可以利用0.0到1.0之间的实数来对应每个特定值属于某个给定类别的程度，因此这里利用模糊逻辑就可以描述“高收入”这样一个模糊概念，而无需使用大于50K的这样一个硬性标准。

在进行分类的数据挖掘系统中，引用模糊逻辑概念，具有在较高的抽象层次上进行数据挖掘的优势，这体现了模糊逻辑在分类中具有非常重要的作用。基于规则系统运用模糊逻辑一般包括以下几个操作。

（1）属性值需转化成模糊值，如图6.11所示，就是将一个连续取值属性income映射到离散类别中（低收入、中等收入和高收入），并计算出相应的模糊值（概念隶属度）。模糊逻辑系统通常都会提供相应操作工具来帮助用户完成这一映射工作。

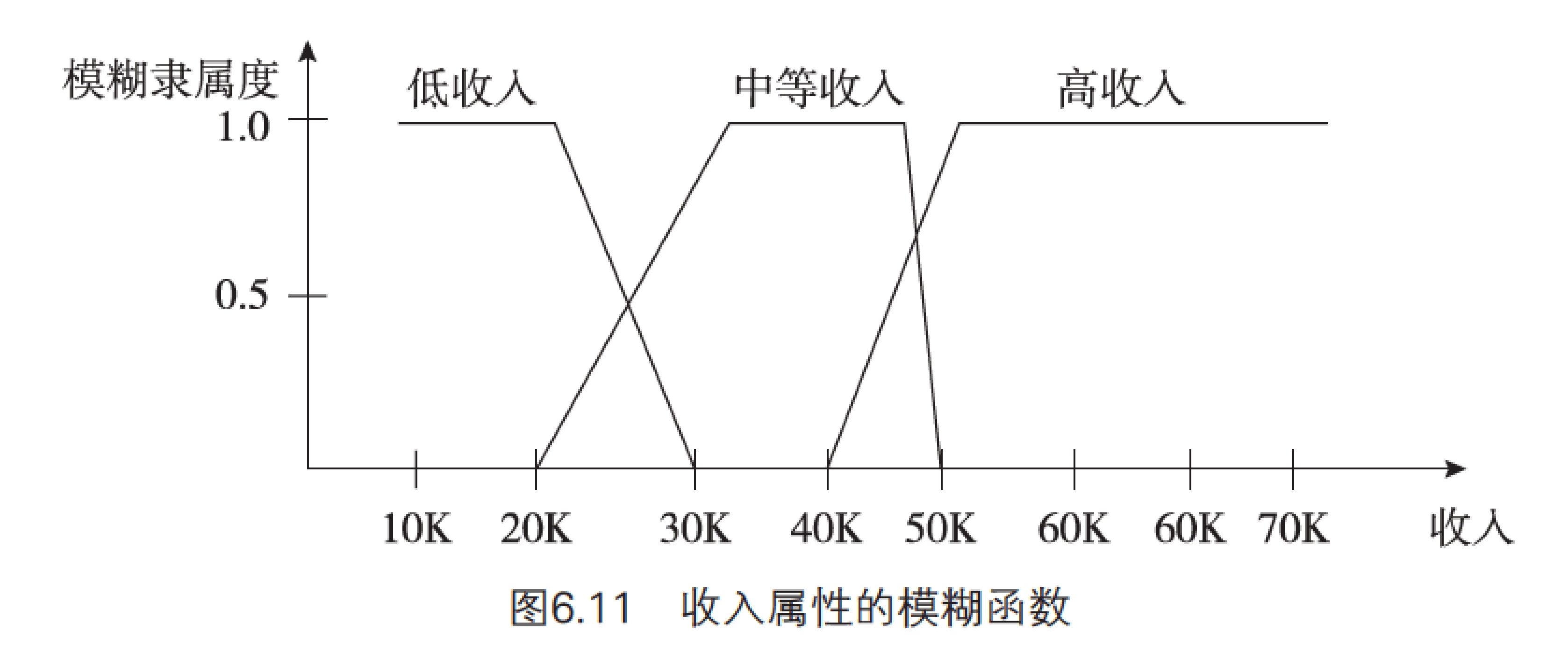


图6.11　收入属性的模糊函数

（2）对于一个给定的新样本，可以应用多个规则。在这些被应用的规则中，每一个规则都贡献一票给概念隶属度的计算。通常情况下，要得到最终的结果，需要累加每个预测类别的相应隶属度，即模糊值。

（3）系统返回在步骤（2）中得到的隶属度总和。可以对每个隶属度增加一个权重，即每个隶属度乘以相应的权重值，然后再进行累加操作。依赖模糊隶属函数是比较复杂的，其计算可能也很复杂。

目前，模糊集分类法已经应用到许多领域中，比如健康医疗和金融保险等领域。

6.3.7　回归

回归分析（regression analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。回归分析运用十分广泛，它按照涉及自变量的多少，可分为一元回归分析和多元回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。如果在回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变

量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。

1. 线性和多元回归

对一个连续数值的预测可以利用统计回归方法所建的模型来实现。线性回归是一种最简单的回归方法，利用一条直线来描述相应的数据模型。二元回归模型的因变量Y可用自变量X的线性函数表示，如下所示：

$$Y=\alpha+\beta X \tag{6-33}$$

其中Y的方差为常数，α 和 β 是回归系数，分别表示直线在Y轴上的截距和直线的斜率。可以运用最小二乘法对这些系数进行求解，使得实际数据与该直线的估计之间误差最小。给定s个样本或形如$(x_1,y_1), (x_2,y_2), \ldots,(x_s,y_s)$的数据点，回归系数 α 和 β 可用如下公式计算：

$$\beta=\frac{\sum_{i=1}^{s}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{s}(x_i-\overline{x})^2} \tag{6-34}$$

$$\alpha=\overline{y}-\beta\overline{x} \tag{6-35}$$

其中$\overline{x}$是$x_1, x_2, \ldots, x_s$的平均值，$\overline{y}$是$y_1, y_2, \ldots, y_s$的平均值。

多元回归是线性回归的扩展，涉及到多个自变量。因变量Y可以是一个多维特征向量的线性函数。基于两个预测属性或变量X_1，X_2的多元回归模型的例子如下。

$$Y=\alpha+\beta_1X_1+\beta_2X_2 \tag{6-36}$$

同理，利用最小二乘法可以获得α，β_1和β_2的数值。

2. 非线性回归

通过在基本的线性回归模型公式中添加高阶项（项的次数大于1），得到多项式的回归模型。一般运用变量转换方法将非线性模型转换为能够应用最小二乘法来求解的线性模型。

现有一个如下公式的三阶多项式：

$$Y=\alpha+\beta_1X+\beta_2X^2+\beta_3X^3 \tag{6-37}$$

为了将其用转换为线性回归模型，可以增加两个新变量，公式如下：

$$X_1=X;\ X_2=X^2;\ X_3=X^3 \tag{6-38}$$

公式可以转换成线性形式，如下所示：

$$Y=\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3 \tag{6-39}$$

这样，利用最小二乘法就可以获得这一公式的各项系数：α、β_1、β_2、β_3。有一些模型本身就是非线性不可分的，对于这些情况，若想得到最小二乘情况下的近似，可能需要通过对更复杂的公式进行计算。

3. 其他回归模型

对于连续取值的函数，可以运用线性回归对其建立模型。线性回归因其较为简单的特点而得到了广泛地应用。对于离散取值变量，可以运用广义线性模型对其进行回归建模。在广义线性模型中，因变量Y的变化速率是自变量X均值的一个函数，而线性回归中因变量Y的变化速率是一个常数，这是两者的一个区别。常见的广义线性模型有对数回归和泊松回归。其

中对数回归模型是以一些事件发生的概率作为自变量的线性回归模型；泊松回归模型主要描述数据出现次数，因为这些数据的出现次数常常表现为泊松分布。

6.4　聚类分析

聚类分析旨在发现紧密相关的观测值群组，使得与属于不同簇的观测值相比，属于同一簇的观测值相互之间尽可能相似。相异度是基于描述对象的属性值来计算的，距离是经常采用的度量方式。聚类分析源于许多研究领域，包括数据挖掘、统计学、生物学以及机器学习。

本节将介绍一些常用的聚类分析方法，以及数据对象间距离的具体计算过程，该计算方法的依据是数据对象的属性。通常聚类方法主要有：划分方法、层次方法、基于模型的方法、基于密度的方法、基于网格的方法和双聚类方法等。

6.4.1　基本概念

聚类是数据挖掘、模式识别等研究中的一个重要内容，在识别数据的内在结构方面具有极其重要的作用。聚类主要应用于模式识别中的语音识别、字符识别等，机器学习中的聚类算法应用于图像分割和机器视觉，图像处理中聚类用于数据压缩和信息检索。聚类的另一个主要应用是数据挖掘（多关系数据挖掘）、时空数据库应用（GIS等）、序列和异类数据分析等。此外，聚类还应用于统计科学，对生物学、心理学、考古学、地质学、地理学以及市场营销等研究也都有重要的作用。

一个聚类的经典定义：一个类簇内的实体是相似的，不同类簇的实体是不相似的；一个类簇是测试空间中点的汇聚，同一类簇的任意两个点间的距离小于不同类簇的任意两个点间的距离；类簇可以描述为一个包含密度相对较高的点集的多维空间中的连通区域，它们借助包含密度相对较低的点集的区域与其他区域（类簇）相分离。

事实上，聚类是一个无监督的分类，它没有任何先验知识可用。

典型的聚类过程主要包括数据（或称之为样本或模式）准备、特征选择和特征提取、接近度计算、聚类（或分组）及对聚类结果进行有效性评估等步骤。

（1）数据准备。包括特征标准化和降维。

（2）特征选择。从最初的特征中选择最有效的特征，并将其存储于向量中。

（3）特征提取。通过对所选择的特征进行转换形成新的突出特征。

（4）聚类。首先选择合适特征类型的某种距离函数（或构造新的距离函数）进行接近程度的度量，而后执行聚类或分组。

（5）聚类结果评估。是指对聚类结果进行评估。

评估主要有3种：外部有效性评估、内部有效性评估和相关性测试评估。

没有任何一种聚类算法可以普遍适用于各种多维数据集所呈现出来的多种多样的结构。通常，需要根据实际的情况选择较为合适的聚类分析算法，比如选择时要考虑应用所涉及的数据类型、聚类的目的、具体应用要求等因素。根据数据在聚类中的积聚规则以及应用这些规则的方法，有多种聚类算法。

6.4.2 划分方法

对于一个给定的包含n个数据对象的数据库，要把其中的对象分成K个聚类，划分方法就是运用一些相关的算法将对象集合划分成k份（$k \leqslant n$），其中每个划分表示一个聚类。较好的聚类划分体现在：属于一个聚类中的对象是“相似”的，而属于不同聚类中的对象是“不相似”的。通常，要求所得到的聚类使得客观划分标准（常称为相似函数，如距离）最优化以达到较好的聚类划分效果。

比较常用的划分方法包括K-Means、k-Medoids、EM算法等，以及一些在这些算法的基础上做了改进的算法。

1. k-Means算法

k-Means算法以k为参数，把n个对象分为k个簇，以使类内具有较高的相似度，而类间的相似度最低。相似度的计算根据一个簇中对象的平均值（被看作簇的重心）来进行。

k-Means算法的处理流程是：首先，随机地选择k个对象，每个对象初始地代表了一个簇中心；对剩余的每个对象，根据其与各个簇中心的距离，将它赋给最近的簇；然后重新计算每个簇的平均值；不断重复这个过程直到准则函数收敛。通常采用均方差作为标准测度函数，定义如下：

$$E=\sum_{i=1}^{k}\sum\nolimits_{p\in C_i}\left|p-m_i\right|^2 \qquad (6\text{-}40)$$

其中E表示数据库中所有对象的均方差总和，p表示空间中代表某个对象的一个点，m_i表示聚类C_i的均值。

这一聚类标准的目的是使得到的k个聚类满足：各个聚类本身具有较高的相似度，而各聚类之间在最大程度上分开。该算法的计算复杂度为O(nkt)，其中n表示对象个数，k表示聚类个数，而t表示循环次数，通常情况下，k<<n和t<<n。该算法常常终止于局部最优。由此，可以看出k-Mean算法在一定程度上可以有效地处理大数据库。

2. k-Medoids算法

k-Means算法存在着一些不足，比如异常数据会影响该算法各聚类均值的计算，影响对数据分布的估计，而且该算法对离群点很敏感。因此，人们在k-Means算法的基础上做了一些改进，得到了k-Medoids算法。不同于k-Means算法运用各聚类的均值作为聚类中心，k-Medoids算法运用medoid作为参考点，然后计算各个对象与各个参考点之间的距离即差异性之和，根据这个总和最小化的原则，应用划分方法来实现对象的聚类划分。

k-Medoids算法的基本操作过程是：首先随意选择一个对象代表每个簇，剩余的对象根据其与代表对象的距离分配给最近的一个簇。然后反复地用非代表对象来替代代表对象，以改进聚类的质量。聚类结果的质量用一个代价函数来估算，即评估对象与其参照对象之间的平均相异度。

3. EM算法

期望最大化（Expectation Maximization，EM）算法不将对象明确地分到某个簇，而是根据

表示隶属可能性的权来分配对象。也就是说，在簇之间没有严格的边界。新的均值基于加权度量值计算。

在实际应用中，有相当多的问题属于数据残缺问题。不能直接观察到的变量称为隐含变量，任何含有隐含变量的模型都可以归为数据残缺问题。EM算法是解决数据残缺问题的一个十分有效的算法。

6.4.3　层次方法

层次聚类方法的基本思路是将数据分为若干组并形成一个组的树从而进行聚类。一般有两种基本层次聚类方法，一种是自下而上聚合层次聚类方法（AGNES），该方法的基本操作为：先将每个对象自身作为一个聚类，然后聚合这些聚类以得到更大的聚类，当所有对象都聚合成为一个聚类，或满足一定终止条件时操作完成。另一种是自顶而下分解层次聚类方法（DIANA），该方法先将全部的对象当成一个聚类，然后不断分解这个聚类以得到更小的聚类，这个过程中小聚类的个数不断增多，当所有对象都独自构成一个聚类，或满足一定终止条件时操作完成。上述两种层次聚类方法的操作过程是相反的。具体过程如图6.12所示。

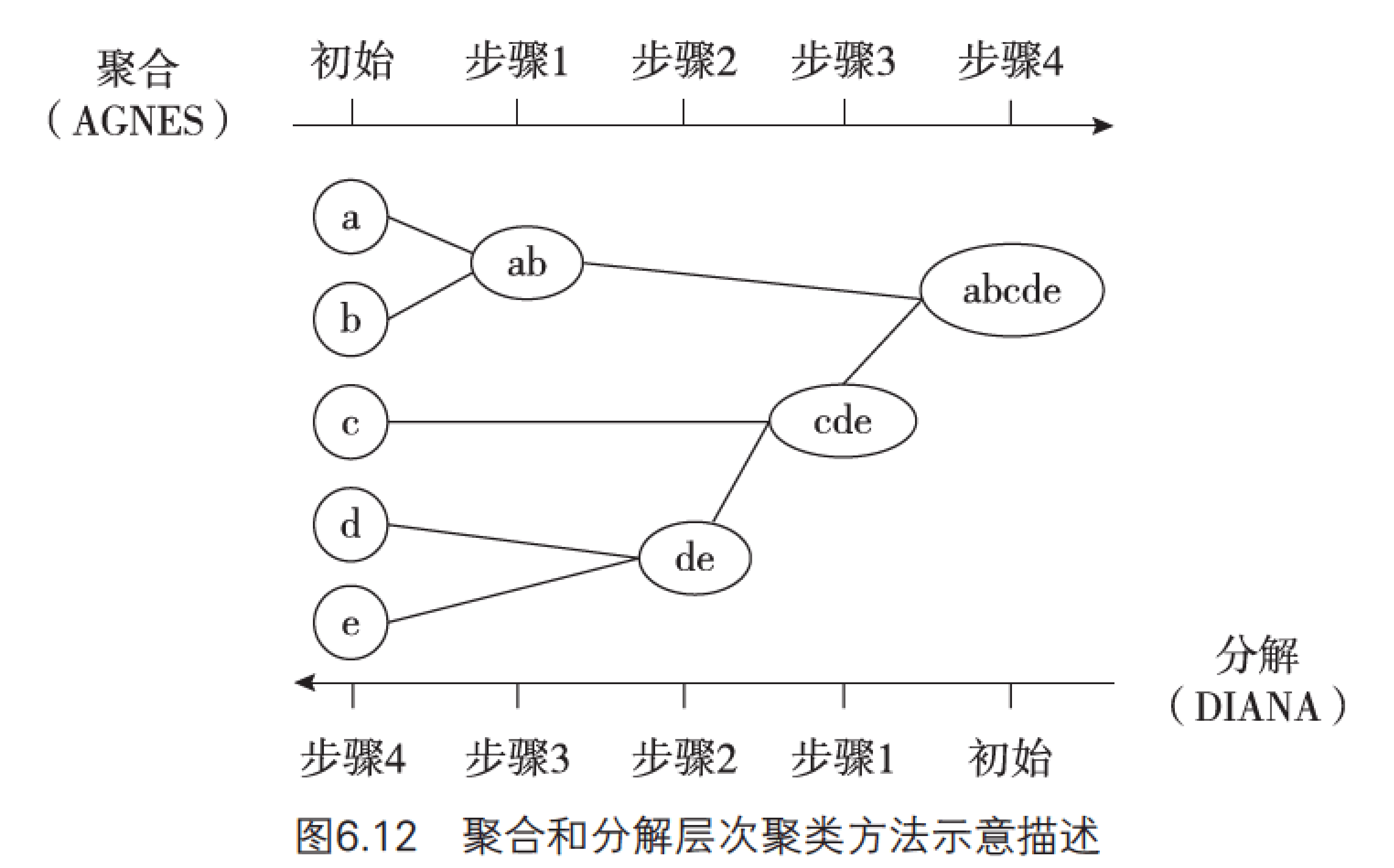


图6.12　聚合和分解层次聚类方法示意描述

层次算法能够产生高质量的聚类，但也存在计算和存储需求较大，缺乏全局目标函数，合并决策不能撤销等问题。因此，经常将层次方法与其他聚类技术相结合以进行多阶段的聚类，比较有著名的方法有BIRCH（Balanced Iterative Reducing and Clustering using Hierarchies）和CURE（Clustering Using Representative）等。

1. BIRCH

BIRCH方法是一种集成的层次聚类方法，聚类特征（Clustering Feature，CF）和聚类特征树（CF Tree）是该方法的两个重要概念。聚类描述可以运用这两个概念来进行概要总结。相应的有关数据结构有利于使聚类方法提高聚类速度，而且也有利于提高处理大数据库的可拓展性。另外，BIRCH方法能够有效地进行增量和动态聚类。

聚类特征是一个三元组，该三元组储存了对象子集的概要信息。设一个子聚类包含N个d–维数据或对象δ_i，则此子聚类的CF定义为：

$$CF=(N,\ \overrightarrow{LS},\ SS) \tag{6-41}$$

其中，N表示该子聚类包含的对象个数，$\overrightarrow{LS}$表示这N个点之和，即$\sum_{i=1}^{N}\overrightarrow{\delta_i}$，$SS$表示数据点的平方和，即$\sum_{i=1}^{N}\overrightarrow{\delta_i^2}$。聚类特征的作用基本上就是总结给定的子聚类的统计信息，其中包括聚类计算以及空间存储利用所需的关键信息。

BIRCH方法操作主要有以下两个阶段。

第一阶段：扫描数据库，建立一个基于内存的初始CF树。此树保留了数据中包含的有关聚类结构的信息，在一定程度上，可以看成对数据的压缩。

第二阶段：选择一个合适的聚类算法，对CF树的叶结点进行聚类。

BIRCH方法采用多阶段处理的方式为：首先扫描一遍数据，从而获得一个基本理想的聚类；通过第二次扫描数据来帮助改善所获聚类的质量。BIRCH的计算复杂度为$O(n)$，其中n为带聚类的对象数。

2. CURE

CURE算法属于聚合方法与分解的中间做法，它描述一个聚类不仅仅是采用一个聚类中心或者对象来进行，而是选取固定数目具有代表性的空间点来进行一个聚类表示。表示聚类代表性的点的产生，先要对分布较好的聚类对象进行选择，接着按照特定的速率（收缩因子）把它们“收缩”或是移至聚类的中心。每一步算法，都是合并拥有分别来源于两个不同聚类两个代表性点所涉及的两个聚类。

每个包含超过一个代表性点的聚类对于CURE方法调整好它的非圆状边界非常有帮助，聚类的收缩或是压缩都对压制异常数据起到一定的作用。因此CURE方法对异常数据表现得更具鲁棒性，同时它也能识别具有非圆形状和不同大小的聚类。此外，该方法在不牺牲聚类质量的情况下，对大数据库的处理也具有较好的可扩展性。

3. ROCK

ROCK也是一个聚合层次聚类算法，与CURE不同，ROCK适合处理符号属性，一般度量两个簇的相似度是根据比较集合的互连性与用户定义的互连性模型来进行的。ROCK先是通过相似度阈值以及共同邻居的概念在给出的数据相似度矩阵中构建一个稀疏的图，接着再在这个图上运行一个层次聚类算法。

4. Chameleon

Chameleon算法是一种在层次聚类中采用动态模型的聚类算法，是在CURE算法与ROCK算法有所不足的基础上提出的。CURE算法与它的相关方案忽略了两个不同簇中数据项的聚集互联性信息；而ROCK算法及其相关方案则忽略了两个不同簇的接近度。Chameleon算法既考虑两个簇之间的互联性，又考虑两个簇之间的接近度。在建立两个不同簇间互联度和接近度时，利用每个簇内部对象的特征来定义相似的子簇。因而，Chameleon并不依靠于静态的、用户提供的模型，只要定义了相似函数就能够应用于各种类型的数据。

6.4.4 基于密度的方法

基于密度方法可以在具有任意形状的聚类的发现上提供帮助。通常在一个数据空间中，低密度（稀疏）的对象区域（一般就认为是噪声数据）将会分割高密度的对象区域。

1. DBSCAN

DBSCAN（Density-Based Spatial Clustering of Applications with Noise）是一种基于密度的聚类算法。这种算法能够把拥有较高密度的区域划分成簇，同时能将空间数据库中任意形状的聚类从“噪音”中分离，它将簇定义为密度相连的点的最大集合。

基于密度的聚类的基本想法和一些相关的定义，具体如下。

（1）一个给定对象的 ε 半径内的近邻就叫做这个对象的 ε -近邻。

（2）若一个对象的 ε -近邻包含超过最低数目（MinPts）个对象，就把这个对象叫做核对象。

（3）给定一组对象集D，若q为核对象，一个对象p为q的 ε -近邻，那么就说p是从q可以“直接密度可达”。

（4）对于一个 ε 来说，一个对象p是从对象q可“密度可达”；一组对象集D有MinPts个对象；如果有一组对象$p_1,p_2,\ldots,p_n$，其中$p_1=q$，$p_n=p$，则使得（对于 ε 和MinPts来讲）p_i+1是从p_i可“直接密度可达”，其中有$P_i \in D$，$1 \leqslant i \leqslant n$。

（5）对于 ε 和MinPts来讲，若存在一个对象$o(o \in D)$，使得从o可“密度可达”对象p与对象q，对象p为“密度连接”对象q。

密度可达为密度连接的一个传递闭包，这种关系是不对称的，仅有核对象是相互“密度可达”，而密度连接是对称的。

DBSCAN检验数据库中每一个点的 ε -近邻。若一个对象p的 ε -近邻包含超过MinPts个，就要创建包含p的新聚类。接着DBSCAN再根据这些核对象，循环收集“直接密度可达”的对象，其中可能涉及合并若干“密度可达”聚类。当各聚类再无新点（对象）加入时聚类进程结束。

2. OPTICS

DBSCAN在进行聚类时需要指定输入参数 ε 和MinPts，但在实际操作时参数很难确定，而且一般算法对参数的设置敏感，参数的变化对聚类结果影响非常大。为了克服这个问题，人们提出了OPTICS（Ordering Points to Identify the Clustering Structure）聚类顺序方法，这种方法不会直接生成一个聚类，而是为自动及交互的聚类分析计算得到一个簇次序。该次序表示了基于密度的数据聚类结构，包括了与基于许多参数设置所得到的基于密度聚类相当的信息。

OPTICS算法给数据库中的对象建立一个对象顺序，同时保存每个对象核心距离与一个适当的可达距离，这些信息足够帮助通过任意小于产生聚类顺序 ε 的距离，产生全部的密度聚类。

6.4.5 基于网格的方法

基于网格聚类方法通过多维网格数据结构，它把空间分成数目有限的单元，以构成一个能够进行聚类分析的网格结构[①]。该方法的最主要的特点就是它处理时间和数据对象数目不

① http://www.docin.com/p-98743940.html

相关，但与每维空间所划分的单元数相关，因此基于网格聚类方法的处理时间很短。

1. STING

STING（Statistical Information Grid）是一个基于网格多分辨率的聚类方法，它把空间分成方形单元，不同层次的分辨率相对的是不同层次的方形单元。这些单元组成了一个层次结构，高层次单元被分解成一组层次相对低些的单元。涉及各网格单元属性的统计信息（如：最小值、最大值、均值）皆可以事先运算和存储。

一个自上而下基于网格方法来处理查询的操作步骤是：首先依据查询内容对层次结构的开始层次进行明确，一般该层次所含的单元较少[①]；对于当前层次中的各个单元，计算其信任度差（或估计概率范围）来反映当前单元和查询要求的相关程度，消除不相关的单元以便仅考虑相关单元，一直重复这一过程直至到达最底层；此时如果满足查询要求，就返回满足要求的相关单元区域，否则取出相关区域单元中的数据，进一步处理它们直至查询要求得到满足。

2. CLIQUE

CLIQUE（CLustering In QUEst）聚类算法是将基于密度以及基于网格的聚类方法综合在一起，它在大型数据库中的高维数据的聚类中起着重要的作用。CLIQUE的中心思想是：给出一个多维数据点的大集合，一般数据点在数据空间中是非均衡分布的。CLIQUE可以将空间中稀疏的和拥挤的区域区别开来，从而发现数据集合的全局分布模式。若一个单元中的包含的数据点多于了某个输入的参数，那么这个单元是密集的。在CLIQUE中，簇就是相连的密集单元的最大集合。

此外，在实际操作中也常常会用到基于模型的聚类方法，主要是统计学方法以及神经网络方法这两种模型，该方法是基于"数据是根据潜在的概率分布生成的"这个假设，试图对给定的数据和某些数学模型之间的适应性进行优化。

6.4.6 基于模型的方法

基于模型的方法是假设每个聚类的模型同时发现与模型相应的模型数据。通常有统计学方法以及神经网络方法等。基于统计学的聚类方法主要分为三种：Cheeseman与Stutz一起提出的AutoClass，Fisher提出的COBWEB以及Gennari等人提出的CLASSIT。

1. 统计学方法

目前在产业界比较流行的一种聚类方法是AutoClass，它通过贝叶斯统计分析来对结果簇的数目加以估算。这个系统采用搜索模型空间存在一切分类的可能性，来对分类类别的个数以及模型描述的复杂性进行自动确定。它允许在一些类别内属性之间可以存在一定的相关性，各个类之间存在一定的继承性，也就是在类层次结构中，一些类共同享一定的模型参数。

COBWEB是一种流行、简单的增量概念聚类算法，它创建层次聚类的形式是一个分类树。分类树和判定树并不一样，通常分类树里面一个节点就对应一个概念，包括这个概念的概率描述，能将这个节点的对象信息加以概括。判定树标记的不是节点而是分支，而且运用逻辑描述符，而不是概率描述符。

---

① http://www.docin.com/p-548850492.html

CLASSIT是在COBWEB的基础上加以扩展，主要对连续性数据的增量聚类进行处理。这种算法在各个节点中存储属性的连续正态分布（也就是均值及标准差），运用修正的分类效用度量，这种度量并非是在离散属性上求和，而是连续属性上的积分。然而CLASSIT存在和COBWEB相似的问题，同样不适合对大型数据库中的数据加以聚类。

2. 神经网络方法

神经网络方法是把一个簇描述成一个样本。而聚类的原型则是样本，无需非得与特定的数据实例及对象相对应。神经网络聚类方法通常包括由Rumelhart等所提出的竞争学习神经网络以及Kohonen所提出的自组织特征映射（SOM）神经网络。采用神经网络聚类方法需要的处理时间较长，并且有较高的数据复杂性。需要研究能够提高网络学习速度的学习算法，并增强网络的可理解性，以便使人工神经网络处理方法适用于大型数据库。

6.4.7 双聚类方法

寻常的聚类是按照数据的全部信息对数据聚类，该方法称之为传统聚类。传统聚类仅仅可以用来寻找全局信息，没办法找出局部信息，而海量的生物学信息却隐藏于局部信息当中。为帮助人们寻找这些信息，2000年，CHENG与CHURCH一起提出了双聚类（bicluster）概念，同时对双聚类进行定义。

双聚类分析方法就是在行和列两个方向上进行聚类分析，通常采用贪婪迭代搜索的方法来发现子矩阵或稳定的类，这些子矩阵中感兴趣的模式具有特定的生物学意义，在很大程度上克服了一些传统聚类分析方法带有的缺陷。为打破传统聚类的局限性以及更好地提高效率，许多算法在寻找双聚类时都采用了贪婪迭代搜索方法。

1. CC算法

CC算法采用逐渐删除能够使子矩阵的平均平方残差降低的行与列，获得一个初步的双聚类，然后逐步增加不会令子矩阵平均平方残差增多的行与列，从而获得一个较好的双聚类。为了找出更多双聚类，算法对已经找到的双聚类进行随机数覆盖，接着删除或添加过程继而获得特定个数的双聚类结果。算法可以较迅速地取得用户指定数目的双聚类。

设X是基因集，Y是对应的表达条件集。a_{ij}是基因表达数据矩阵A中的元素。设I、J分别是X、Y的子集，那么（I，J）对指定的子矩阵A_{IJ}具有下面的平均平方残基：

$$H(I,J)=\frac{1}{|I||J|}\sum_{i\in I,j\in J}\left(D_{ij}-D_{iJ}-D_{Ij}+D_{IJ}\right)^2 \qquad (6\text{–}42)$$

其中：$a_{iJ}=\frac{1}{|J|}\sum_{j\in J}a_{ij}$和$a_{Ij}=\frac{1}{|I|}\sum_{i\in I}a_{ij}$分别为行平均值、列平均值和子矩阵（$I$，$J$）的平均值。对于δ ≤0，则称该子矩阵为一个δ –bicluster。

为了高效地寻找δ –bicluster，作者使用了一种增删节点的方法来寻找均方残差最小的子矩阵。首先，使用多节点删除法，逐渐删除能够降低子矩阵的平均平方残差的行与列，获得一个初步的双聚类；然后，使用单节点删除法，精细删除某一行和列，只要操作后矩阵的均方残差比原来的均方残差的减小，直至无法下降；最后，只要操作后矩阵的均方残差比原来的均方残差的减小，使用节点插入法，精细增加某一行和列，直至无法下降。由于CC算法是

一种确定性算法，每次都是得到一个相同的双聚类，每一个节点皆与数据矩阵中的一行或者一列相对应。

CC算法实现了对基因表达数据在基因和条件两个方向的同时聚类，减弱了相似度算法对聚类的影响，采用平均平方残基得分方案来评价结果矩阵的质量，提高了聚类准确性。然而这样也有着很大的缺陷，随机数的替换会造成原始数据的更改，导致结果的不精确，也没办法找出重叠的双聚类，同时还容易陷进局部最优的缺陷。Yang等人根据CC算法加以改进，进而提出了FLOC算法。

2. FLOC算法

FLOC算法的打分原则类似于CC算法，不同的是CC算法考虑的是所有都是确定值，而FLOC算法则在缺失值上面有着独特的处理。该算法第一步是生成一定数量的种子，接着采用计算对某一行或列进行添加或删除，每一步都尽力让双聚类的中间结果增益的变化达到最大。尽管FLOC算法能够找出可重叠的双聚类，但是在很大程度上双聚类结果的好坏和运行时间都对初始聚类有着非常大的依赖，而这些初始聚类的产生一般都是随机的。双聚类的贪心策略效率比较高，然而聚类结果容易陷进局部最优。为弥补贪心策略会陷进局部最优的缺陷，某些算法先是运用贪心策略来寻找双聚类，接着再应用智能优化算法处理找到的双聚类从而得到比较好的结果。例如Ste-Fan等人就改良了CC算法，即在添加或删除过程当中好的行列其保留概率较大，反之则较小，迭代所获的结果即是种子，运用进化算法优化来得到较好的双聚类。

假设一个基因表达矩阵$A=(a_{ij})_{max}$具有n个基因m个条件，X是这个矩阵的基因全集，Y是这个矩阵的条件全集。有一个双向聚类（I，J），其中$I\subseteq X,J\subseteq Y$，那么对于在双向聚类里面的全部元素$a_{ij}$，$i\subseteq I, j\subseteq J$，有以下规定：

$$a_{ij}=\frac{1}{|J_i'|}\sum\nolimits_{j\in J_i'}a_{ij} \tag{6-43}$$

其中J'_i是第i行中的确定值（非缺失值）。

$$a_{ij}=\frac{1}{|I_i'|}\sum\nolimits_{i\in I_i'}a_{ij} \tag{6-44}$$

其中I'_i是第j行中的确定值（非缺失值）。

$$a_{ij}=\frac{1}{|V_{ij}|}\sum\nolimits_{i\in I,\, j\in J}a_{ij} \tag{6-45}$$

这里的a_{ij}是表示的是双向聚类（I，J）中确定值的元素，而$|V_{ij}|$叫做双向聚类（I，J）的容量，也就是双向聚类（I，J）中确定值的元素的个数。

为了寻找k个容量较大，剩余值较小的双聚类，作者首先运用统一概率参数p生成k个容量为$(M\times p)\times(N\times p)$，且容量大于容量阈值的双聚类种子循环；然后，生成一个操作序列（①固定序列，②随机序列，③随机权重序列），顺序执行操作序列，依次选择最优操作，生成一个（$m+n$）个双聚类组，判断是否进入下个循环，还是结束退出；最后，更新初始双聚类组，用（$m+n$）个双聚类组中平均子矩阵剩余值r最小的双聚类组代替初始双聚类组。假如产生操作序列，则需要采用交换概率跑$p(i+j)$修改操作序列里的次序。

FLOC算法能够一次得到多个双向聚类，运用跳过的方法来处理缺失值，在某种程度上

保障了数据矩阵的稳定性，双向聚类质量在多次迭代的作用下显得的更好。但是因为FLOC算法是多次迭代，并且碍于迭代条件的缘故，其运行速度较慢，迭代过程中寻优的两个目标分离，最终导致所寻找的目标无法明确。

表6.3所示为FLOC算法与CC算法的比较。

表6.3　FLOC算法与CC算法比较表

	CC算法	FLOC算法
双聚类类型	连贯值型	连贯值型
双聚类结构	任意位置重叠	任意位置重叠
双聚类结果	一次一个	一次多个
算法分类	贪婪迭代搜索	贪婪迭代搜索
算法目标	在阈值范围内的最大双聚类	剩余值尽可能小容量尽可能大
缺失值的处理	随机数替换	跳过不处理

CC算法是最为经典的算法，很多算法都是把它当作航标。即便这样，CC算法依然存在很多需要改进的地方，FLOC算法采取两方面的处理打破了CC算法的一些局限。就整体来说，这两种算法依然存在非常多的相似性质，主要表现在双向聚类的结构、类型和算法分类上面。当然两者还是有区别的，这从双向聚类的结果、算法目标与对缺失值的处理局能够轻易发现。FLOC算法也可以看作CC算法的改进算法，但是还存在其不足或是需要改进的地方。

6.5　离群点检测

离群点检测的任务是识别特征显著不同于其他数据的观测值，这样的点称为异常点、离群点或孤立点。离群点检测算法的目标是发现真正的离群点，同时避免将正确的对象标注为离群点。离群点检测的应用包括欺诈检测、入侵检测、故障检测、疾病的不寻常模式、生态系统扰动等。

本节主要介绍离群点检测的基本概念，以及离群点检测常用的一些算法，包括基于统计的、基于距离的和基于偏差的检测算法。

6.5.1　基本概念

在不同领域，离群点有不同的定义，通常情况下描述离群点检测可以定义为：给定一个有n个数据点或对象的数据集和期望的离群点数目k，找出与数据集中其余数据显著不同的、异常的或不一致的前k个对象。离群点挖掘问题可以被看做两个子问题：（1）定义在给定的数据集中，什么样的数据可以被认为是不一致的；（2）找到一个有效的方法来挖掘所定义的离群点。

导致离群的原因主要包括：（1）数据来源于异类，如欺诈、入侵、疾病爆发、不寻常的实验结果等。（2）由数据变量固有变化引起，是自然发生的，反映了数据集的数据分布特征，如气候变化、顾客的新的购买模式、基因突变等。（3）数据测量和收集误差，主要是由于人为错误、测量设备故障或存在噪音。根据不同的分类标准，可以将离群点分成不同的类别，如图6.13所示。

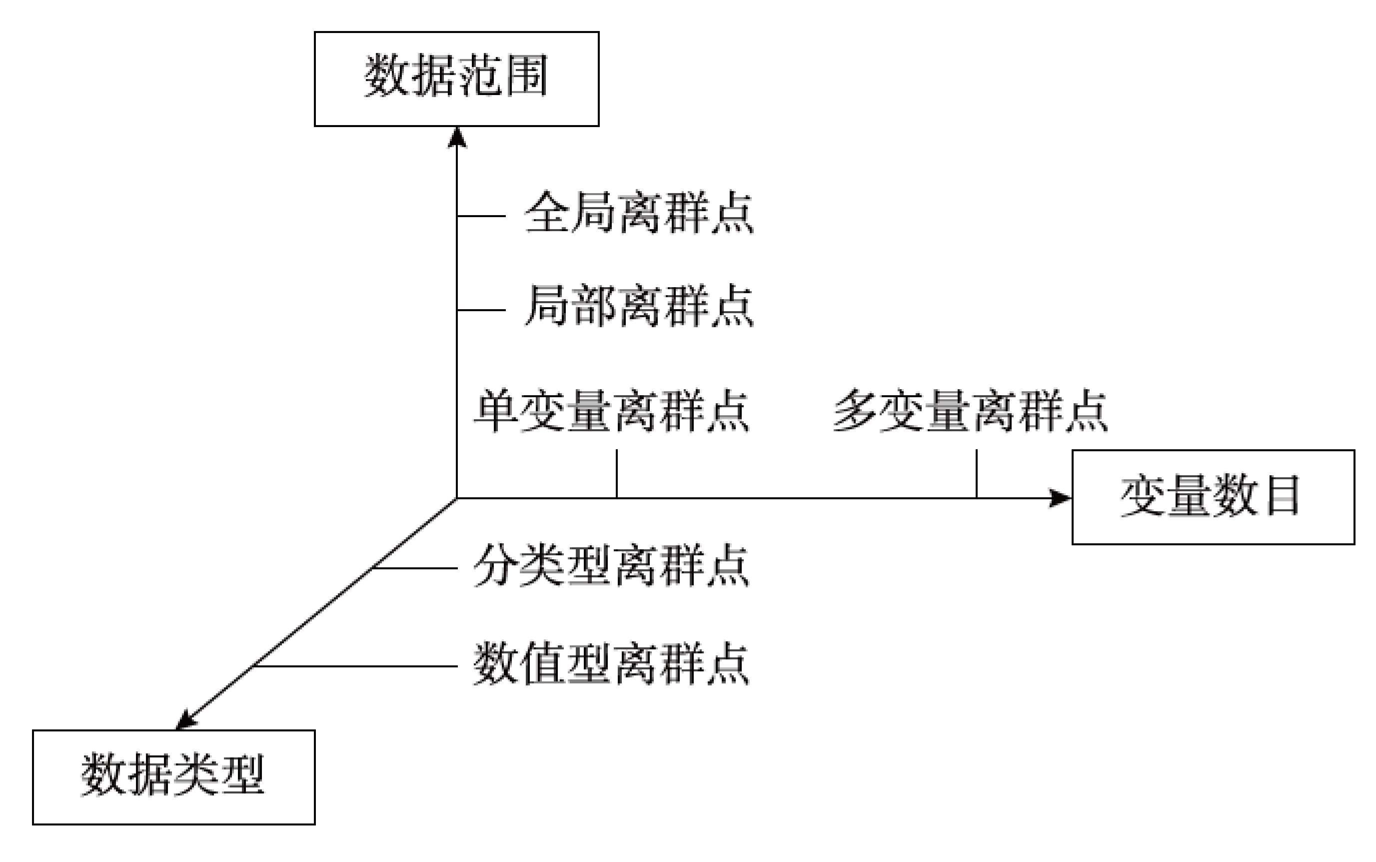


图6.13　离群点分类

6.5.2　基于统计的离群点检测

统计方法首先对已知给定的数据样本集假设一个分布或者概率模型（比如正态分布或泊松分布），然后采用不一致性检验，结合假设的模型，从而确定离群点。这种检测需要事先了解数据集的有关参数（如数据分布情况）、分布的参数（如均值和方差）以及期望的离群点数目。

一个统计不一致测试检查两个假设，即一个正面假设和反面假设。一个正面假设H就是一个描述n个数据对象来自一个分布F，即：

$$H: o_i \in F, \text{其中} i=1,2,\ldots,n \qquad (6\text{-}46)$$

若没有非常明显的统计证据来反驳H，则接受H，认为H成立。一个不一致测试就是验证一个对象O_i与分布F关系是否非常大（或小）。根据不同的数据知识，提出了不同的统计量，得到了不同的不一致测试方法。假设选择统计量T作不一致性测试，对象o_i的统计值是v_i；构造分布T并对重要性概率SP（v_i）$=Pr\ ob$（$T>v_i$）进行评估。若存在某个概率SP（v_i）足够小，则认为对象o_i是不一致的，拒绝正面假设。一个反面假设$\overline{H}$，是一个描述对象o_i来自另一个分布G。分布F模型的选择对判断对象是否一致具有非常大的影响，是因为对象o_i在一个模型是不一致数据，但是在另外一个模型中可能就是一致的。此外，反面模型对对象o_i的判断也是非常重要的，即确定当对象o_i确实是不一致数据时正面假设被拒绝的概率。

运用统计方法进行离群点检测存在的主要不足：因为检测出多维空间中的不一致数据是大多数数据挖掘问题所需要的，但是离群点统计检验测试基本上都是针对单个属性的，这在很大程度上限制了此方法的应用范围。此外，运用统计方法还需要知道数据集参数的有关知识，比如数据的分布情况。然而在很多情况下，是不知道数据的分布情况的。而且统计方法也不能保证可以检测出所有的不一致数据，特别是当不采用特殊的具有针对性的测试方法，或者任何标准分布都不能描述数据时，要检测出不一致数据是不容易实现的。

6.5.3　基于距离的离群点检测

基于距离的离群点检测方法是针对运用统计方法进行不一致数据的检测所存在的不足提出来的。对象o是一个基于距离的孤立点，通常用DB(p,d)表示，其中p和d为对象o的参数，它

代表的意义是数据集合S中至少存在p部分对象与对象o的距离大于d。也就是说，独立于统计检验，将那些与给定对象的距离较大的对象看作是基于距离的孤立点。与运用统计的方法进行离群点检测相比较，基于距离的不一致数据检测综合归纳了基于标准分布模型的不一致性检验。基于距离的不一致性数据检测在一定程度上降低了计算量，这些计算通常是因为检测方法的选择和标准分布的拟合等操作产生的。

对于很多异常值检验，如果一个对象o对于一个给定的检验是一个不一致数据，那么对恰当的参数p和d，对象o也是一个DB(p,d)不一致数据。目前，已经有不少基于距离的离群点检测方法被提出，例如基于索引的算法、嵌套循环算法和基于单元算法等。基于距离的离群点检测需要手动设置参数p和d，而找到合适的参数需要不断地尝试，这个过程中可能出现错误的参数设置。

6.5.4 基于偏差的离群点检测

基于偏差的离群点检测是一种通过检测对象的特征从而找出不一致性数据的方法。如果一个对象偏离了给定对象的特征描述，那么认为此对象是不一致的。也就是说该方法中的“偏差”指的是异常。

基于偏差的离群点检测方法主要有以下两种。

1. 序列异常技术

该技术模仿了人类可以从一系列类似的对象中识别出异常对象的行为，它使用隐含的数据冗余。给定一个包含n个对象的数据集D，建立一个子集合序列$\{D_1, D_2,\cdots,D_m\}$，$2\leqslant m\leqslant n$，并有$D_{j-1}\subset D_j$；$D_j\subseteq D$，评估序列中子集之间的差异度。这个技术介绍了如下几个关键术语。

（1）异常集。就是偏离的集或离群点，它被定义为对象的最小子集，这个子集的移除导致数据集中剩余部分的差异度有最大量的减少。

（2）差异度函数。这个函数不要求对象间的一个度量距离，它可以是任何函数，只要满足条件，即给定一个对象集，如果某个对象与另一个对象是相似的，那么就返回一个较低的值。对象之间的差异度越大，函数返回值也越大。

（3）平滑因子。对于序列中每个子集，这个函数都会被计算，它估价了从原始对象集中移除子集后，差异度能减少多少，这个值通过集的基数来测量。平滑因子最大的子集就是序列中的异常集。

这个算法选择一个子集的序列来进行分析，对于每个子集，它确定这个子集与序列中前一个子集的差异度。为了避免输入顺序对结果产生任何可能的影响，处理过程可以重复若干次，每一次都有子集的一个不同的随机顺序。在所有的迭代中，有最大平滑因子值的子集就是异常集。

2. OLAP数据立方体技术

一个进行偏离检测的OLAP方法使用数据立方体来辨识高维数据中的异常区域。为了提高效率，偏离检测过程可以与立方体计算过程重叠进行。该方法是一种探索驱动的方法，预先计算的只是数据异常的度量，被用来在数据集合计算的所有层次上指导用户进行数据分析。

如果数据立方体中的一个单元值与基于一个统计模型的期望值显著不同，那么这个单元值就被认为是一个异常。这个方法采用如背景颜色那样的可视化提示来反映每个单元的异常程度，用户可以选择对标志为异常的单元进行向下钻取。一个单元的度量值可能反映发生在该立方体上的更细节或更低层次上的异常，因为这些异常在当前层次上可能是不可见的。

这个模型考虑了一个单元所属的所有维上的度量值的变化和模式。例如，假设有一个销售数据的数据立方体，并且要查看每个月的销售概况。在可视化提示的帮助下，注意到与其他月相比，十一月的销售量有所增加，在时间维上这像是一个异常。可是，通过下载十一月来观察这个月的每一项的销售，可以注意到十一月中其他项的销售也有一个相似的增加。因此，如果考虑到项目维，十一月的总销售上的增加就不是一个异常。这个模型考虑了隐藏在所有数据立方体集合分组操作后面的异常。因为搜索的空间往往是比较大的，尤其当维数较大且涉及到多层时，要发现异常，仅仅依靠人工手动进行检测是难以实现的。

6.6 复杂数据类型挖掘

数据挖掘涉及到多学科以及多领域，正是由于数据挖掘越来越广泛的应用，需要处理的数据类型也多种多样，复杂数据类型挖掘成为数据挖掘中一项重要的前沿研究课题。本节将介绍处理一些复杂数据类型的挖掘方法，主要包括文本挖掘、Web挖掘、时空数据挖掘和多媒体数据挖掘等。

1. 文本挖掘

随着信息技术的发展，互联网数据及资源呈现海量特征，传统的信息检索技术已无法适应日益增加的大量文本数据处理的需求，人们提出了文本挖掘的方法，并将不同的文档进行比较，并进行文档重要性和相关性排列，找出多文档的模式或趋势。

文本挖掘的一般处理过程包括了对大量文档集合的内容采取预处理、特征提取、结构分析、文本摘要、文本分类、文本聚类、文本关联等手段[①]。文本挖掘不仅要处理大量的结构化和非结构化的文档数据，而且还要处理其中复杂的语义关系，因此，现有的数据挖掘技术无法直接用于文本挖掘。关于非结构化数据的挖掘，目前有两种思路：一种是开发新的数据挖掘算法，直接展开非结构化数据的挖掘，因为数据类型的复杂性，使得这种算法的复杂程度非常的高；另一种是把非结构化数据变成结构化的数据，再通过已有的数据挖掘技术展开挖掘工作。现在的文本挖掘技术通常都采用这种方法进行。而对于语义关系，就需要集成计算语言学以及自然语言处理等方面的成果来加以分析。

2. Web挖掘

可以说互联网是一个分布广泛、全球性的庞大的信息服务中心，它提供了多种多样的信息服务，诸如教育、新闻、广告、政府、金融管理、消费信息、电子商务等。Web中含有极

---

① http://wenku.baidu.com/link?url=cJsn4fWgjKZ77xnu9y8PcqzQyAjn77aNgXfWiudc02cEOArrNADRyw9RzgZyuF1s71KhcJMHDVUhXMsVJ6fVcwe9D7J7oN1oe4hR5G63LWK.

其丰富与动态的超链接信息，同时还包括对Web页面的访问与使用信息，这些都给数据挖掘提供了丰富的资源。

可以把Web挖掘定义成：从与WWW有关联的资源及行为中获取有用的模式以及隐含的信息。通常Web挖掘可划分成3类：（1）在文档内容或是它的描述中获取知识的过程，即Web内容挖掘；（2）从WWW的组织结构以及链接关系中推导的知识，即Web结构挖掘；（3）从Web的访问记录中抽取出感兴趣的模式，即Web使用记录的挖掘。

3. 时空数据挖掘

20世纪90年代中后期，数据挖掘领域的一些较成熟的技术，如关联规则挖掘、分类、预测与聚类等被逐渐用于时间序列数据挖掘和空间数据挖掘，以发现与时间或空间相关的有价值的模式。随着传感器网络、全球定位系统（GPS）、手持移动设备和射频识别（RFID）等设备的普遍应用，这些设备产生并积累了大量的移动对象数据。此外，遥感卫星和地理信息系统（GIS）等技术的显著进步，使人们前所未有地获取了大量的气候数据、数字影像数据以及地理科学数据。这些时空数据内嵌于连续空间，其样本在时间、空间上存在很强的自相关性，其中隐含的模式往往是局部的，从而使时空数据挖掘具有特殊性和复杂性。

按照数据挖掘的定义，可将时空数据挖掘定义为从具有海量、高维、高噪声和非线性等特性的时空数据中提取出隐含的、人们事先不知道的、但又潜在有用的信息及知识的过程。时间维和空间维为数据挖掘任务增加了额外的复杂性。按照挖掘的任务主要可分为以下几类：时空模式发现、时空聚类、时空异常检测、时空预测和分类等。

4. 多媒体数据挖掘

随着网络的快速发展以及信息技术的不断进步，数据形式也日益多样化，在这样的环境下，涌现出大量的多媒体数据。通常多媒体数据类型有图像、视频、音频、时空数据以及超文本等。这些多媒体数据隐藏着大量潜在有价值的知识。多媒体数据挖掘是综合分析海量的多媒体数据的视听特性和语义，挖掘出隐藏在其中的、具有一定价值的、能够理解的知识模式，找出事件的趋势以及关联性，从而为用户的决策提供参考。

对数据的挖掘以及对多媒体信息的处理是多媒体信息挖掘的两个重要方向。研究多媒体信息挖掘会遇到很多挑战，比如针对媒体数据的特性，如何才能有效地运用数据挖掘的理论和方法进行多媒体数据挖掘；利用多媒体的时间、空间、视觉特性、视听对象以及运动特性等内容，如何快速有效地挖掘出具有价值的知识规律等。

6.7　数据挖掘的研究前沿和发展趋势

数据挖掘涉及多个学科领域，它融合了统计学、数据库、数据仓库、机器学习、人工智能、模式识别、可视化等技术的研究成果，数据挖掘应用的领域非常广泛。只要数据是有分析价值的，都可以应用数据挖掘工具挖掘出有用的信息。数据挖掘典型的应用领域包括金融、医疗、零售和电商、电信、交通等。另外，由于新的数据类型也随着技术进步不断增加，因此本节还指出了数据挖掘的发展趋势和所面临的挑战。

6.7.1 数据挖掘的应用

数据挖掘所要处理的问题，就是在庞大的数据中找出有价值的隐藏事件，并加以分析，获取有意义的信息和模式，为决策提供依据。数据挖掘应用的领域非常广泛，只要有分析价值与需求的数据，都可以利用挖掘工具进行发掘分析。目前，数据挖掘应用最集中的领域包括金融、医疗、零售和电商、电信和交通等，而且每个领域都有特定的应用问题和应用背景。

1. 金融领域

不管是银行还是其他金融机构都存储了海量的金融数据，比如信贷、储蓄与投资等金融数据。对于这些数据，运用数据挖掘技术进行有针对性地处理，将会得到很多具有价值的知识。金融数据具有可靠性、完整性和高质量等特点，这在很大程度上利于开展数据挖掘工作以及挖掘技术的应用。数据挖掘在金融领域中有许多具体的应用，例如，分析多维数据，以把握金融市场的变化趋势；运用孤立点分析等方法，研究洗黑钱等犯罪活动；应用分类技术，对顾客信用进行分类，为维持与客户的关系以及为客户提供相关服务等决策提供参考。

2. 医疗领域

在人类的遗传密码、遗传史、疾病史以及医疗方法等医疗领域中，都隐藏着海量的数据信息。另外，对医院内部结构、医药器具、病人档案以及其他资料等的管理也产生了巨量的数据。对于这些巨量的数据，运用数据挖掘相关技术进行处理，从而得到相关知识规律，将有利于相关人员工作的开展。运用数据挖掘技术，在很大程度上有助于医疗人员发现疾病的一些规律，从而提高诊断的准确率和治疗的有效性，不断促进人类健康医疗事业的发展。

3. 零售和电商领域

由于零售业会产生庞大的数据，主要是销售数据，比如商品的购进卖出记录、客户购买、消费记录等。特别是随着在Web以及电子商务等商业方式日益普及流行，相应地数据也以飞快的速度增长着。运用数据挖掘技术对这些海量的数据进行针对性的处理分析，可以获取很多极具价值的知识。例如可以有效地识别顾客的购买行为，从而把握好顾客购买的趋势。这些关于顾客的有效信息是商家采取最佳决策的关键依据。商家可以根据数据挖掘结果有针对性的采取有效措施，比如如何改进服务质量，确保顾客的满意度；如何提高商品的销售量；如何设计较优的运输路线以及采取怎样的销售策略等，从而提高企业效益。此外，由于数据挖掘的推荐系统已经成为电子商务的关键技术，通过数据挖掘再对网站进行系统分析，对用户的行为模式加以识别，在增加客户粘性，提供个性化服务，优化网站设计等方面也取得了很好的效果。

4. 电信领域

电信运营商已逐渐发展为一个融合了语音、图像、视频等增值服务的全方位立体化的综合电信服务商。三网融合，即电信网、因特网和有线电视网的“融合”，是未来的一种发展趋势，这一现象将会产生巨量的数据。运营商要合理的分析商业形式和模式，运用数据挖掘是非常有必要的。例如：对用户行为、利润率、通信速率和容量、系统负载等电信数据，可以运用多维分析方法进行分析；要发现异常模式，可以运用聚类或孤立点分析等方法进行数

据挖掘；要得到电信发展的影响因素，可以运用关联或序列等模式进行分析等。总之，数据挖掘技术对电信业的发展发挥着非常重要的作用，比如如何提高相关资源的利用率、更深入更充分地了解用户行为、如何获取更多的经济效益等。

5. 交通领域

交通问题对城市的民生有很大影响，该领域积累了大量的数据比如出租公司积累的乘客出行数据和公交公司的运营数据。通过对乘客数据和运营数据进行分析和挖掘，能够为公交、出租公司科学的运营和交通部门的决策提供依据，比如合理规划公交线路，实时为出租车的行驶线路提供建议等。这样，不但可以提升城市运力和幸福指数，还能有效减少因交通拥堵问题造成的成本浪费。另外，航空公司也可依据历史记录来寻找乘客的旅行模式，以便提供更加个性化的服务，合理设置航线等。

近年来，数据挖掘的应用发展迅速，不仅在以上领域，在政府部门、军事、制造业、科学研究等方面也都取得了一定的进展。

6.7.2　数据挖掘中的隐私问题

隐私权是指个体的私人信息不被他人非法收集、公开和利用的权利。隐私保护就是保护个体的隐私权不被侵害，保护个体隐私在未经授权的时候不被泄露和恶意利用。基于隐私的数据挖掘存在两个层面的问题：

（1）原始信息隐私保护。企业、医院、政府部门通常收集了大量的个人原始信息，泄露这些信息可能识别出个人用户的身份。为了防止个人隐私的泄露，这些原始数据均需要在进行数据挖掘之前进行修改和隐藏。这个层面主要解决的问题是如何在原始数据不准确的前提下得到正确的挖掘结果。

（2）敏感规则隐私保护。企业、医院、政府部门不仅存储着大量的个人原始信息，通过对这些原始信息的挖掘，还可以得知某一群体的特征和行为规律。为了防止这些敏感规则被挖掘出来，通常事先改变原始数据的统计特征，使这些敏感规则的生成概率大大降低。

我们既不能否认通过数据挖掘产生的巨大利益，也不能因为存在有隐私保护的问题就废弃数据挖掘，而是应当正视存在的隐私保护的现状和方法。目前隐私保护技术正得到越来越多的关注，在保护隐私信息方面还需要更多的探索。更好的一个愿景是，从计算机科学、管理科学、社交网络技术、政策法规等多个方面有效的结合在一起，共同来完成从数据中安全和无泄漏地发现有效的知识。

6.7.3　数据挖掘的发展趋势

数据挖掘已慢慢地从高端的研究转向日常的应用，在金融业、零售业等一些对数据分析需求比较大的领域已经成功地采用了数据挖掘技术来辅助决策。尽管如此，由于技术的进步和社会的发展，数据挖掘技术仍然面临着许多新的问题和挑战。

1. 数据挖掘与物联网、云计算和大数据

简单来说，物联网就是物物相连的网络，是数字世界与物理世界的高度融合。物联网底

层的大量传感器为信息的获取提供了一种新的方式，这些传感器不断地产生着新的数据，随着各种各样的异构终端设备的接入，物联网采集的数据量也就会越来越大，其数据类型和数据格式也会越来越复杂。这些数据与时间和空间相关联，有着动态、异构和分布的特性，也为数据挖掘任务带来了新的挑战。

云计算是一种基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态、易扩展且经常是虚拟化的资源（包括硬件、平台和软件），实现了设备之间的数据应用和共享。随着物联网的发展，感知的信息不断增加，需要不断地增加服务器的数目来满足需求，但由于服务器的承载能力是有限的，使得服务器在节点上出现混乱和错误的几率大大增加。为了更好地提供服务，基于云计算的系统能有效地解决物联网分布式数据挖掘中所遇到的问题，在进行相关数据挖掘时能够显著的提高性能。

目前，大数据已成为继物联网、云计算之后又一信息科技的新热点。大数据在本质上仍然是海量数据，但规模更大，实时性和多样性特点更明显，相应地数据挖掘技术也需要有所改进，研究如何处理半结构化甚至非结构化的数据是目前大数据挖掘面临的挑战之一。

将物联网、云计算、大数据与数据挖掘研究联系起来，不仅具有深远的科学研究价值，而且将产生巨大的经济效益和社会价值。

2. 数据挖掘研究和应用面临的挑战

大数据时代的数据挖掘面临着新的挑战，主要表现在以下几个方面。

数据类型的多样性：不同的应用、系统和终端，由于标准的差异性，会产生不同结构的数据，其中包括结构化数据，半结构化数据和非结构化数据，对这些异构化数据的抽取与集成将成为一大挑战。

数据挖掘算法的改进：大数据时代数据的量级达到了一个新的阶段，而且还有其他新的特征，现有挖掘算法需要基于云计算进行改进，以适应不同应用对数据处理能力的需求。

数据噪声太大：由于普适终端的所处地理位置的复杂性，使得产生的数据具有很多噪声。在进行数据清洗时，不易把握清洗粒度。粒度太大，残留的噪声会干扰有价值的信息；粒度太小，可能会遗失有价值的信息。

数据的安全性与隐私保护：互联网的交互性，使得人们在不同地点产生的数据足迹得到积累和关联，从而增加了隐私暴露的概率，且这种隐性的数据暴露往往是无法控制和预知的。随着数据挖掘工具和电子产品的日益普及，保护隐私和信息安全是数据挖掘将要面对的一个重要问题。这就需要进一步地开发，以便在适当的信息访问和挖掘过程中保护隐私与信息安全。

3. 数据挖掘的发展方向

应用的探索：数据挖掘正在探索、扩大其应用范围，通用数据挖掘技术在处理特定应用时存在着局限性。因此，目前有一种针对特定应用来开发数据挖掘系统的趋势。

可视化数据挖掘：可视化能更直观地展示数据的特性，具图像展示更符合人的观察习惯。可视化数据挖掘已成为从大量数据中发现知识的有效途径，系统研究和开发可视化数据挖掘技术将推进数据挖掘作为数据分析的基本工具。

数据挖掘与数据库/数据仓库系统和其他应用系统的集成：数据库/数据仓库系统等已经

成为信息处理系统的主流，而且与数据库和数据仓库系统的紧耦合方式正是数据挖掘系统的理想体系结构。将不同的系统集成到统一的框架中，有利于保证数据的可获得性和一致性，以及数据挖掘系统的可移植性、可伸缩性和高性能。

数据挖掘的应用在很多领域取得了一定的成果，而且其广阔的应用前景已吸引了众多的研究人员和商业公司的加入。但是数据挖掘所带来的有关隐私和信息安全的问题，需要着重考虑。数据挖掘技术发展的时间很短，属新兴科学，在技术和社会不断发展的今天，还面临着很多挑战和值得重点研究的方向，相信数据挖掘技术的研究与应用将会得到长足的进步，必将产生巨大的经济和社会效益。

6.8　练习

1. 简述数据库知识发现的过程。

2. 介绍数据挖掘任务有哪几种类型，各种类型的含义是什么。

3. 给出一个未在本章讨论的关联分析的案例。

4. 简述分类的基本思想和解决分类问题的一般过程，并举例说明如何利用分类方法预测用户购买电脑的模型。

5. 简述聚类分析的原理和过程，说明k-Means算法的基本思想和聚类过程。

6. 介绍离群点检测的类型，并介绍一下离群点检测有哪些应用场景。

7. 归纳数据挖掘面临的挑战和发展的趋势。

参考文献

[1] 徐义明. 天文光谱分类算法在分布式环境下的应用研究[D]. 济南：山东大学，2008.

[2] 罗可. 数据库中数据挖掘理论方法及应用研究[D]. 长沙：湖南大学，2004.

[3] 张金. 数据挖掘技术在3G业务扩展中的研究与应用[D]. 长沙：湖南师范大学，2010.

[4] 牛文颖. 改进的ID3决策树分类算法在成绩分析中的应用研究[D]. 大连：大连交通大学，2008.

[5] 杨萍. 决策树和关联规则在教学评价系统中的应用[D]. 北京：北京工业大学，2013.

[6] 罗华群. 校园一卡通数据的挖掘与应用[D]. 上海：华东师范大学，2009.

[7] 刘兴林. 数据挖掘技术在银行业中的应用[D]. 重庆：重庆大学，2005.

[8] 王越. 分布式关联规则挖掘的方法研究[D]. 重庆：重庆大学，2003.

[9] 罗可，林睦纲，郗东妹. 数据挖掘中分类算法综述[J]. 计算机工程，2005，31（1）：3–5.

[10] 刘红岩，陈剑，陈国青. 数据挖掘中的数据分类算法综述[J]. 北京：清华大学学报（自然科学版），2002，42（6）：727–730.

[11] 刘亮晴. 基于人工神经网络的施工招标评标系统研究[D]. 重庆：重庆大学，2006.

[12] 张跃强. 基于神经网络的结构损伤检测方法研究[D]. 北京：北京工业大学，2000.

[13] 郑浩. 基于人工神经网络的高层建筑结构选型[D]. 泉州：华侨大学，2000.

[14] 齐金鹏. 数据挖掘模型可视化研究及其应用实例[D]. 长春：吉林大学，2004.

[15] 国刚. 基于数据挖掘的客户忠诚度分析[D]. 青岛：青岛大学，2004.

[16] 张远春. 多品种混合装配车间关键资源建模与优化[D]. 上海：上海交通大学，2011.

[17] 钟世刚. 面向CRM的贝叶斯分类算法及并行化研究[D]. 重庆：重庆大学，2003.

[18] 刘利民. 基于粗糙集的分类规则挖掘的研究[D]. 辽宁：辽宁工程技术大学，2004.

[19] [美]Jiawei Ham等著. 数据挖掘：概念与技术（原书第3版）[M]. 范明，孟晓峰（译）. 北京：机械工业出版社，2012.

[20] 黄珺. 晶圆失效图形自动识别系统在ULSI良率管理中的应用研究[D]. 上海：复旦大学，2009.

[21] Pang-Ning Tan,Michael Steinbach等著. 数据挖掘导论（完整版）[M]. ，范明，范宏建（译）. 北京：人民邮电出版社，2011.

[22] 郭军华. 数据挖掘中聚类分析的研究[D]. 武汉：武汉理工大学，2003.

[23] 李雄飞，李军. 数据挖掘与知识发现[M]. 北京：高等教育出版社，2003.

[24] 张敏，戈文航. 双聚类的研究与进展[J]. 微型机与应用，2012，4（4）：4-6.

[25] 刘伟. 基于FLOC的双向聚类算法研究[D]. 广州：华南理工大学，2011.

[26] 徐翔，刘建伟，罗雄麟. 离群点挖掘研究[J]. 计算机应用研究，2009，26（1）：34-40 .

[27] 王宏威. 油田数据挖掘技术的研究与应用[D]. 大庆：大庆石油学院，2005.

[28] 宋艳. CRM中基于CABOSFV改进算法的客户聚类研究[D]. 哈尔滨：哈尔滨工程大学，2004.

[29] 刘大有，陈慧灵，齐红等. 时空数据挖掘研究进展[J]. 计算机研究与发展，2013，50（2）：225-239.

[30] 郭群. 多媒体信息挖掘综述[J]. 信息系统工程. 2010，（08）：103.

[31] 李明江，唐颖，周力军. 数据挖掘技术及应用[J]. 中国新通信. 2012，（22）：66-67.

[32] 徐振龙，郭崇慧. 隐私保护数据挖掘研究的简要综述[C].《第七届（2012）中国管理学年会商务智能分会场论文集（选编）》. 2012.

[33] 王惠中，彭安群. 数据挖掘研究现状及发展趋势[J]. 工矿自动化. 2011，2（2）：29-32.

[34] 张春华，王阳. 数据挖掘技术，应用及发展趋势[J]. 现代情报. 2003，23（4）：47-48.

第7章 数据分析语言R

在海量数据爆炸的时代，传统的技术架构已无法满足海量数据处理的需求。陆续出现的新技术与手段给解决海量数据的存储以及数据查询的的延时问题等提供了可能。其中，开源统计分析语言R被广泛应用于统计、金融、医学等多个行业。R的灵活性以及开放性，使得越来越多的数据分析师关注这款软件，其在大数据处理中的应用也越来越被学界和业界所重视。

7.1 R概述

7.1.1 R是什么

R是一套由数据操作、计算和图形展示功能经过整合而形成的套件，是一个开放（GPL）的统计编程环境。它包括[①]：能够有效的进行数据存储和处理的功能，为数据分析和展示提供了强大的图形功能，体系完整的数据分析工具，一套与S-PLUS所基于的S语言一样简单、完善、有效的编程语言以及一套完整的数组（特别是矩阵）计算操作符。前面所说的“环境”（environment）在这里是为了说明R的定位是一个完善、统一的系统，而非像其他统计分析软件那样作为一个专门、不灵活的附属工具。R很适合用于与发展中的新方法所进行的交互式数据分析[②]。

R是诞生于1980年左右的S语言的一个分支，而S语言是由AT&T贝尔实验室开发的一种用来进行数据探索、统计分析和作图的解释型语言。最初S语言的实现版本主要是S-PLUS，S-PLUS是一个商业软件，它基于S语言，并由MathSoft公司的统计科学部进一步完善。R的使用与S-PLUS有很多类似之处，这两种语言有一定的兼容性。S-PLUS的使用手册，只要稍加修改就可作为R的使用手册。R语言是Auckland大学统计系的Robert Gentleman和RossIhaka于1995年开始编制，并且首次发布，后来经过R-1.0（2000年），R-2.0（2004年），R-3.0（2013年）版本。目前R由志愿者组成的国际团队来维持。

R的特点包括以下几方面：

（1）多数商业统计软件都是收费的，但R是一个免费的统计分析软件（环境）。

（2）R是一个内嵌了许多实用的统计分析函数的全面的统计研究平台，其分析结果能直接显示，其中，一些中间结果不但可以在专门的文件中保存，也可以进行下一步分析。在R

① http://wenku.baidu.com/view/2a00afc805087632311212b8.html

② http://wenku.baidu.com/view/00aacaea6294dd88d0d26b7f.html

中可以完成各种类型的数据分析工作。

（3）R拥有超高水准的制图功能，拥有一系列最强大最全面的将复杂数据可视化的功能，其内嵌的作图函数可将产生的图片保存为多种格式的文件（例如jpg，ps，png，pdf，bmp，xfig，emf，pictex），并且能在一个独立的窗口中展示。

（4）帮助功能很完善，R嵌入了一个非常实用的帮助系统，随软件所附的pdf或html帮助文件可以随时通过主菜单打开、浏览或打印。通过help命令可随时了解R所提供的各类函数的使用方法和范例。

（5）可从多个不同数据源获取数据并将数据转化为可用的形式。R可以轻松地从各种类型的数据源导入数据，包括文本文件、数据库管理系统、统计软件，乃至专门的数据仓库，它同样可以将数据输出并写入到这些系统中。

（6）R可运行于多种平台之上，包括Linux，Windows和Mac OS等操作系统。

R语言不断发展和完善的根本原因在于其开放性和灵活性的特点，以及业界最广泛的支持。同时R语言也获得了越来越多业界及学术界的认可。目前，越来越多的高科技企业、金融机构、制药公司以及几乎所有的西方大学与研究机构都在使用R语言来开展数据分析工作。2012年R语言在“过去十二个月中你在实际项目中使用的数据挖掘或分析工具”的调查中，击败了2010年排名第一的Excel和2011年排名第一的Rapidminer，荣登榜首。而KDNuggets在2013年做的“使用何种编程或统计类语言进行分析和数据挖掘”的调查中，R语言击败了Python、SQL、JAVA和SAS，以60.9%的得票率再一次荣登榜首。

7.1.2 R的获取与安装

R的安装文件及说明的获取非常方便，可以在CRAN（Comprehensive R Archive Network）上免费下载，网址为http://cran.r-project.org[①]。由于具有Linux、Mac OS和Windows相应编译好的二进制版本，R的安装非常容易，运行从网站上下载的程序，如R-3.1.0-win32.exe（R for WindowsSetup），按照Windows的提示安装即可。

安装完成后，程序会创建R程序组并在桌面创建R主程序的快捷方式（可以在安装过程中选择不创建），通过快捷方式运行R，便可调出R的主窗口，如图7.1所示。

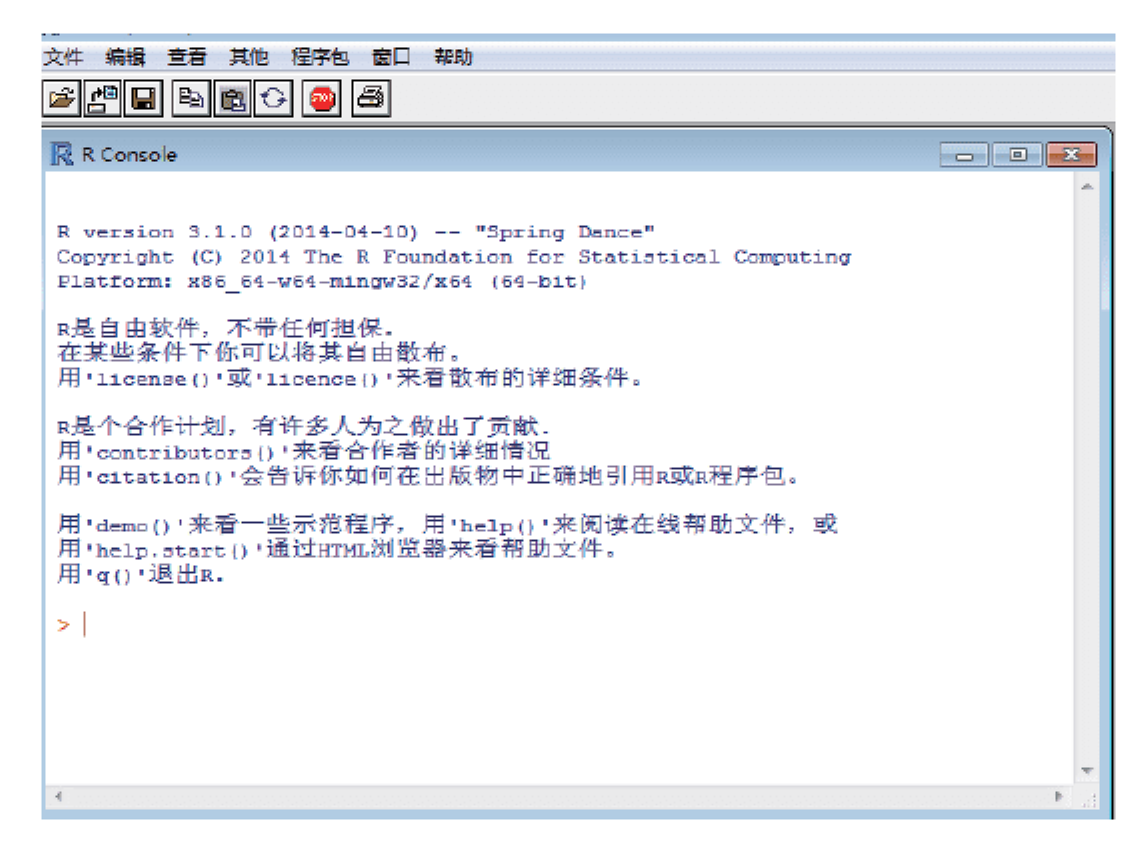


图7.1 R软件主窗口

R软件的操作界面与Windows的其他编程软件相类似，由一些菜单和按钮组成，在菜单下便是命令输入窗口，也是部分运算结果的输出窗口，有些运算结果会在新建窗口中输出。在命令输入窗口中显示的“>”符号是R命令的提示符，当在其后面输入命令，按下回车键后便会给出输出结果。

R最初是一种基于数学脚本的语言，前身为S语言。但是论数学，远不及专业的Matlab和

① http://wenku.baidu.com/view/00aacaea6294dd88d0d26b7f.html

SAS，论脚本功能，又不及Python和Perl。随着R语言的发展，在功能强大的IDE，Rstudio和R本身开源的双重帮助下，R语言成为了一门真正的语言。现在的R语言是面向对象的，可方便调试，可并行处理和接口处理的，程序设计功能也不亚于C/C++、Java等高级语言。

7.1.3　R的使用

1. 新手入门

R是一种区分大小写的解释型语言。可以在命令提示符（>）后每次输入并执行一条命令，也可以一次性执行写在脚本文件中的一组命令。R包括多种数据类型：向量、矩阵、数据框（与数据集类似）以及列表（各种对象的集合）。

R中的多数功能是由程序内置的函数和用户自定义的函数来提供的。一次交互式会话期间的所有数据对象都被保存在内存中。一些基本函数默认是直接可用的，而其他高级函数则包含于按需加载的程序包中。R语句由函数和赋值构成。R使用<-，而不是传统的=作为赋值符号。例如，一个名为n的对象，其内容是数值10，可执行下述命令：

```
>n<-10
> n
[1] 10
```

方括号中的数字1表示从n的第一个元素开始显示。

2. 获取帮助

R提供了大量的帮助文档，学会如何使用这些帮助文档可以在相当程度上助力用户的编程工作。R内置的帮助系统提供了当前已安装包中所有函数的细节、参考文献以及使用示例。帮助文档可以通过表7.1中列出的函数进行查看。

表7.1　R中的帮助函数

函　数	功　能
help.start()	打开帮助文档首页
help("foo")或?foo	查看函数foo的帮助（引号可以省略）
help.search("foo")或??foo	以foo为关键词搜索本地帮助文档
example("foo")	函数foo的使用示例（引号可以省略）
RSiteSearch("foo")	以foo为关键词搜索在线文档和邮件列表存档
apropos("foo", mode="function")	列出名称中含有foo的所有可用函数
data()	列出当前已加载包中所含的所有可用示例数据集
vignette()	列出当前已安装包中所有可用的vignette文档
vignette("foo")	为主题foo显示指定的vignette文档

执行函数help.start()时会打开一个浏览器窗口，用户可在其中查看入门和高级的帮助手册、常见问题集，以及参考材料。函数RSiteSearch()可在在线帮助手册和R-Help邮件列表的讨论存档中搜索指定主题，并在浏览器中返回结果。由函数vignette()函数返回的vignette文档一般是PDF格式的实用介绍性文章。不过，并非所有的包都提供了vignette文档。不难发现，R提供了大量的帮助功能，学会如何使用这些帮助文档，毫无疑问地会有助于编程。笔者经常

会使用？来查看某些函数的功能（如选项或返回值）。

3. 工作空间

工作空间就是当前R的工作环境，它储存着所有用户定义的对象（向量、矩阵、函数、数据框、列表）。在一个R会话结束时，可以将当前工作空间保存到一个镜像中，并在下次启动R时自动载入它。各种命令可在R命令行中交互式地输入。使用上下方向键查看已输入命令的历史记录。这样用户就可以选择一个之前输入过的命令并适当修改，最后按回车键重新执行它。

当前的工作目录是R用来读取文件和保存结果的默认目录。查看当前工作目录的函数是getwd()，当前的工作目录还可使用函数setwd()进行设定。若需要读入一个不在当前工作目录下的文件，在调用语句中应该要写明完整的路径，目录名和文件名并使用引号进行闭合。

用于管理工作空间的部分标准命令见表7.2。

表7.2　用于管理R工作空间的函数

函 数	功 能
getwd()	显示当前的工作目录
setwd("mydirectory")	修改当前的工作目录为mydirectory
ls()	列出当前工作空间中的对象
rm(objectlist)	删除一个或多个对象
help(options)	显示可用选项的说明
options()	显示或设置当前选项
history(#)	显示最近使用过的#个命令（默认值为25）
savehistory("myfile")	保存命令历史到文件myfile中（默认值为.Rhistory）
loadhistory("myfile")	载入一个命令历史文件（默认值为.Rhistory）
save.image("myfile")	保存工作空间到文件myfile中（默认值为.RData）
save(objectlist,file="myfile")	保存指定对象到一个文件中
load("myfile")	读取一个工作空间到当前会话中（默认值为.RData）
q()	退出R，将会询问你是否保存工作空间

首先，当前工作目录被设置为C:/myprojects/project1，当前的选项设置情况将显示出来，而数字将被格式化，显示为具有小数点后三位有效数字的格式。然后，我们创建了一个包含20个均匀分布随机变量的向量，生成了此数据的摘要统计量和直方图。最后，命令的历史记录保存到文件.Rhistory中，工作空间（包含向量x）保存到文件.RData中，会话结束。注意setwd()命令的路径中使用了正斜杠。R将反斜杠（\）作为一个转义符。即使在Windows平台上运行R，在路径中也要使用正斜杠。同时函数setwd()不会自动创建一个不存在的目录。如果必要的话，可以使用函数dir.create()来创建新目录，然后使用setwd()目录指向这个新目录。在独立的目录中保存项目是一个好主意。笔者通常会在启动一个R会话时使用setwd()命令指定到某一个项目的路径，后接不加选项的load()命令。这样做可以让用户从上一次会话结束的地方重新开始，并保证各个项目之间的数据和设置互不干扰。在Windows和Mac OS X平台上完成类似的操作就更简单了，只需跳转到项目所在目录并双击保存的镜像文件即可完成启动R，载

入保存的工作空间，并设置当前工作目录到这个文件夹中。

4. 输入与输出

启动R后将默认开始一个交互式的会话，从键盘接受输入并在屏幕上输出。不过也可以将处理的命令集写在一个脚本文件（一个包含了R语句的文件）中并直接将结果输出到多类目标中。

（1）输入

函数source("filename")可在当前会话中执行一个脚本，当文件名中如果没有包含路径时，R将默认此脚本在当前工作目录中。举例来说，source("myscript.R")将执行包含在文件myscript.R中的R语句集合。依照惯例，脚本文件以.R作为扩展名，不过这并不是必需的。

（2）文本输出

函数sink("filename")将输出重定向到文件filename中。默认情况下，如果文件已经存在，则它的内容将被新内容覆盖。使用参数append=TRUE可以将输出的文本追加到文件的后部，而不会覆盖原有的内容。参数split=TRUE可将输出内容同时发送到屏幕和输出文件中。不加参数调用命令sink()，将仅向屏幕输出结果。

（3）图形输出

虽然sink()可以重定向文本输出，但它对图形输出没有影响。要重定向图形输出，使用表7.3中列出的函数即可。最后使用dev.off()将输出返回到终端。

表7.3　用于保存图形输出的函数

函　数	输　出
pdf("filename.pdf")	PDF文件
win.metafile("filename.wmf")	Windows图元文件
png("filename.png")	PBG文件
jpeg("filename.jpg")	JPEG文件
bmp("filename.bmp")	BMP文件
postscript("filename.ps")	PostScript文件

7.1.4　R包

R提供了大量开箱即用的功能，但它最激动人心的一部分功能是通过可选模块来实现的，这些模块需要下载和安装。目前共有4000多个称为包（package）的用户贡献模块可从http://cran.r-project.org/web/packages下载。这些包提供了横跨各种领域、数量惊人的新功能，比如分析地理数据、处理蛋白质质谱，甚至是心理测验分析的功能。

1. 什么是包

包是R函数、数据、预编译代码以一种定义完善的格式组成的集合。计算机上存储包的目录称为库（library）。函数.libPaths()能够显示库所在的位置，函数library()则可以显示库中有哪些包。R自带了一系列默认包（包括base、datasets、utils、grDevices、graphics、stats以及methods），它们提供了种类繁多的默认函数和数据集。其他包可通过下载后来进行安装。这

些包安装好以后，它们必须被载入到会话中才能使用。查看哪些包已加载并可使用，使用命令search()即可。

2. 包的安装

R中有许多函数可以用来管理包，例如第一次安装一个包，可使用命令install.packages()。不加参数执行命令install.packages()后，将显示一个CRAN镜像站点的列表，在连网的情况下选择其中一个镜像站点之后，将看到所有可用包的列表，选择其中的一个包即可进行下载和安装。如果知道自己想安装的包的名称，可以直接将包名作为参数提供给这个函数。例如，包gclus中提供了创建增强型散点图的函数，可以使用命令install.packages("gclus")来下载和安装它。一个包仅需安装一次。但和其他软件类似，包经常会被其作者更新，使用命令update.packages()可以更新已经安装的包。要查看已安装包的描述，可以使用installed.packages()命令，该命令可以列出安装的包，以及包的版本号、依赖关系等信息。

3. 包的载入

包的安装是指从某个CRAN镜像站点下载它并将其放入库中的过程。要在R会话中使用它，还需要使用library()命令载入这个包。例如，要使用gclus包，执行命令library(gclus)即可。当然，在载入一个包之前必须已经安装了这个包。在一个会话中，包只需载入一次。如果需要，用户可以自定义启动环境以自动载入会频繁使用的那些包。

4. 包的使用方法

载入一个包之后，就可以使用一系列新的函数和数据集了。包中往往提供了演示性的小型数据集和示例代码，能够让用户尝试这些新功能。帮助系统包含了每个函数的一个描述（同时带有示例），每个数据集的信息也被包括其中。命令help(package="package_name")可以输出某个包的简短描述以及包中的函数名称和数据集名称的列表。使用函数help()可以查看其中任意函数或数据集的更多细节。这些信息也能以PDF帮助手册的形式从CRAN下载。

7.2 R的数据操作

R作为一款分析统计软件，擅长对数据进行分析处理，在数据爆炸的时代显得尤为重要。下面，介绍R的数据结构以及数据的输入方法。

7.2.1 数据结构

公认的经典手册《R语言经典入门》（R For Beginners）对R的表示数据的对象描述为[①]：R的运行需要借助于一些对象，一方面是借助于对象的内容和名称，另一方面是对象的数据类型即属性。对象的属性对于作用于一个对象的函数的表现非常重要，正是这个属性为对象提供了所需要的信息。长度和类型是所有对象都具备的两个内在属性。其中长度指的是对象中元素的数目；类型指的是对象中元素的基本类型，包括字符型，数值型，复数型和逻辑型

① http://www.docin.com/p-43353783.html

四种。除此之外还有其他不能用来表示数据的类型，如函数或表达式，其中对象的分别通过函数length和mode得到工作长度和类型[①]。例如执行下面的命令，并观察相应的输出。

```
> x <- 1
> mode(x)
[1] "numeric"
> length(x)
[1] 1
> A <- "Gomphotherium"; compar <- TRUE; z <- 1i
> mode(A); mode(compar); mode(z)
[1] "character"
[1] "logical"
[1] "complex"
```

不管数据的类型是哪一种，总是用NA（不可用）来表示缺失数据。可用指数形式来表示很大的数值：

```
> N <- 2.1e23
> N
[1] 2.1e+23
```

R可以表示无穷的数值，如用Inf和-Inf表示±∞，或者用NaN（非数字）表示不是数字的值。

```
> x <- 5/0
> x
[1] Inf
> exp(x)
[1] Inf
> exp(-x)
[1] 0
> x - x
[1] NaN
```

当输入对象属性为字符型的值时须在值的两端加上双引号。如果在值中出现需要引用双引号的情况时，在引用的双引号的前面须加上反斜杠（\），这两个合在一起的字符（\"）在某些函数如cat的输出显示或write.table写入磁盘函数的qmethod选项时会按照特殊的方式进行处理。例如执行下述的代码：

```
> x <- "Double quotes \" delimitate R's strings."
> x
[1] "Double quotes \" delimitate R's strings."
> cat(x)
```

① http://blog.csdn.net/yangxudong/article/details/7315923

```
Double quotes " delimitate R's strings.
```

用单引号（'）来界定变量是另一种表示字符型变量的方法，此时双引号不需要用反斜杠来引用（不过引用单引号时必须要用！）

```
> x <- 'Double quotes " delimitate R\'s strings.'
> x
[1] "Double quotes \" delimitate R's strings."
```

R拥有许多用于存储数据的对象类型，包括向量、数组、标量、矩阵、数据框和列表。它们在存储数据的类型、创建方式、结构复杂度，以及用于定位和访问其中个别元素的标记等方面均有所不同。

表7.4给出了表示数据的对象的类别概览。

表7.4　数据对象的类别概述

对象	类型	是否允许同一个对象中有多种类型?
向量	数值型、字符型、复数型、逻辑型	否
数组	数值型、字符型、复数型、逻辑型	否
因子	数值型或 字符型	否
矩阵	数值型、字符型、复数型、逻辑型	否
时间序列（ts）	数值型、字符型、复数型、逻辑型	否
数据框 列表	数值型、字符型、复数型、逻辑型 数值型、字符型、复数型、逻辑型	是 是
	函数、表达式，...	

描述数据时，对于一个向量，有类型和长度就可以了，但是对于其它的对象就没这么简单了，需要一些由外在的属性给出的额外信息。在这些属性中dim表示对象维数，如一个2行2列的矩阵，它的dim是一对数值[2,2]，长度是4。而表7.4所列的数据的对象中，向量是一个变量；数组是一个k维的数据表；因子是一个分类变量；而矩阵的维数k=2，是数组的一个特例；时间序列数据包含一些例如频率和时间等的额外的属性，用"ts"来表示；数据框必须是等长的，但可以是不同的数据类型，它是由一个或几个因子和（或）向量构成；列表可以包含任何类型的对象，包括列表。需要特别指出的是，数组或矩阵中的所有元素的类型必须是同一种的[①]。

7.2.2　数据输入

作为一名数据分析人员，通常会面对来自多种数据源和数据格式的数据，这时的任务是将这些数据导入到使用的工具中，然后分析数据，并汇报分析结果。R提供了适用范围广泛的数据导入工具。向R中导入数据的权威指南可在http://cran.r-project.org/doc/manuals/R-data.pdf下载其中的R Data Import/Export手册。R可从键盘、文本文件、Microsoft Excel和Access、流行的统计软件、特殊格式的文件，以及多种关系型数据库中导入数据，如图7.2所示。下面将对这些数据输入方式进行介绍。

① http://blog.csdn.net/yangxudong/article/details/7315923

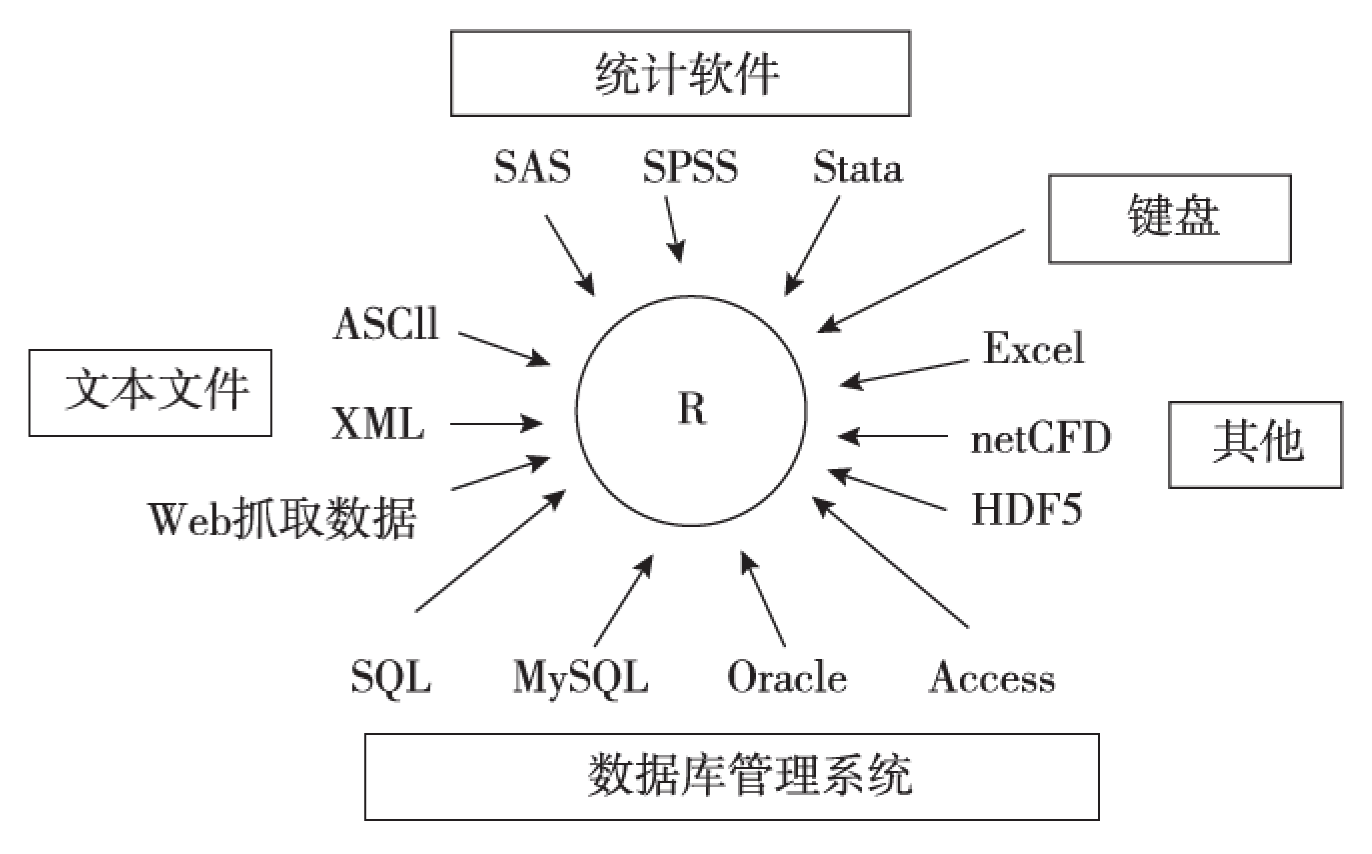


图7.2　可供R导入的数据源

1. 使用键盘输入数据

也许输入数据最简单的方式就是使用键盘了。R中的函数edit()会自动调用一个允许手动输入数据的文本编辑器。具体步骤如下：

（1）创建一个空数据框（或矩阵），其中变量名和变量的模式需与理想中的最终数据集一致。

（2）针对这个数据对象调用文本编辑器，输入需要的数据，并将结果保存回此数据对象中。在下例中，将创建一个名为mydata的数据框，它含有三个变量：age（数值型）、gender（字符型）和weight（数值型）。然后将调用文本编辑器，键入数据，最后保存输入的结果。类似于age=numeric(0)的赋值语句将创建一个指定模式但不含实际数据的变量。注意，编辑的结果需要赋值回对象本身。函数edit()事实上是在对象的一个副本上进行操作的。如果不将其赋值到一个目标，所有修改将会全部丢失。在Windows上调用函数edit()的结果如图7.3所示。

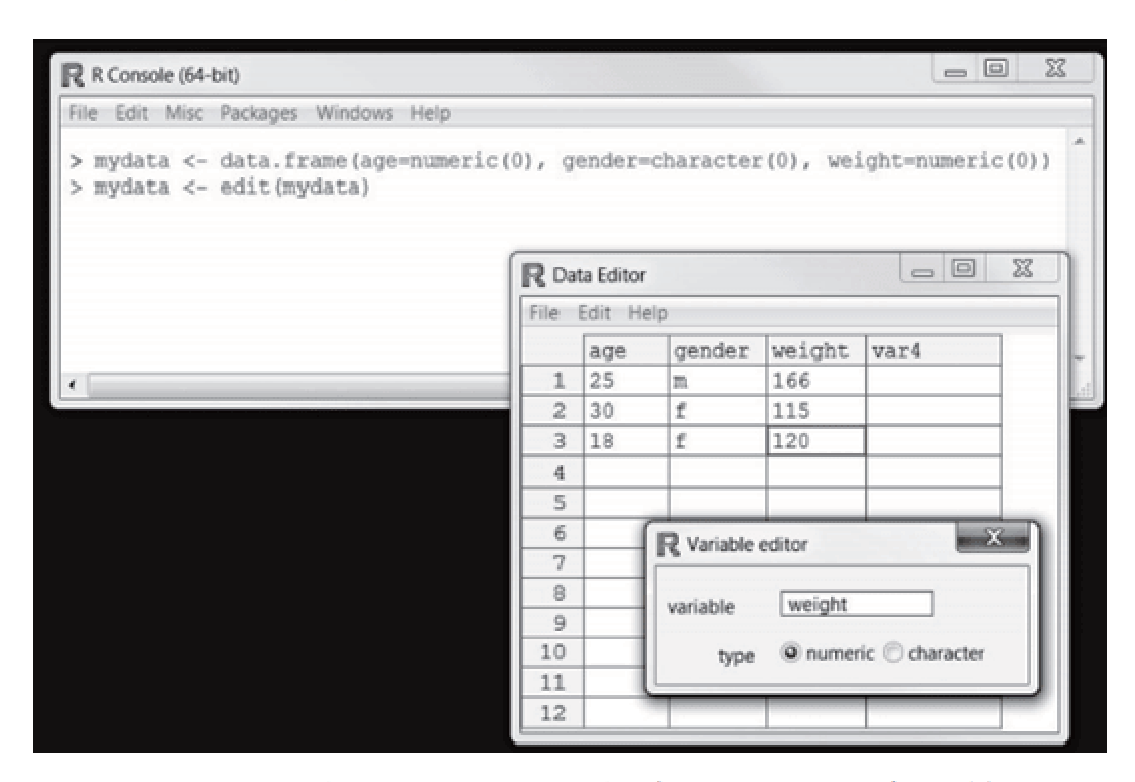


图7.3　通过Windows上内建的编辑器输入数据

如图7.3所示，笔者已经添加了一些数据。单击列的标题，就可以用编辑器修改变量名和变量类型（数值型、字符型）。还可以通过单击未使用列的标题来添加新的变量。编辑器关闭后，结果会保存到之前赋值的对象中（本例中为mydata）。再次调用mydata <– edit(mydata)，就能够编辑已经输入的数据并添加新的数据。语句mydata <– edit(mydata)的一种简捷的等价写法是fix(mydata)。这种输入数据的方式对于小数据集很有效。对于较大的数据集，用户所期望的也许是笔者接下来要介绍的方式：从现有的文本文件、Excel电子表格、统计软件或数据库中导入数据。

2. 从带分隔符的文本文件中导入数据

可以使用read.table()从带分隔符的文本文件中导入数据。此函数可读入一个表格格式的文

件并将其保存为一个数据框。其语法如下：

```
mydataframe<-read.table(file,header=logical_value,sep="delimiter","row.
name=names")
```

其中，file是一个带分隔符的ASCII文本文件，header是一个表明首行是否包含了变量名的逻辑值（TRUE或FALSE），sep用来指定分隔数据的分隔符，row.names是一个可选参数，用以指定一个或多个表示行标识符的变量。举个例子，语句：

```
>grades<-read.table("studentgrades.csv",header=TRUE,sep=",",Row.
name="STUDENTID")
```

从当前工作目录中读入了一个名为studentgrades.csv的逗号分隔文件，从文件的第一行取得各变量的名称，将变量STUDENTID指定为行标识符，最后将结果保存到名为grades的数据框中。

请注意，参数sep允许导入那些使用逗号以外的符号来分隔行内数据的文件。可以使用sep="\t"读取以制表符分隔的文件。此参数的默认值为sep=""，即表示分隔符可为一个或多个空格、制表符、换行符或回车符。在默认情况下，字符型变量将转换为因子。我们并不总是希望程序这样做（例如处理一个含有被调查者评论的变量时）。有许多方法可以禁止这种转换行为。其中包括设置选项stringsAsFactors=FALSE，这将停止对所有字符型变量的此种转换。另一种方法是使用选项colClasses为每一列指定一个类，例如logical（逻辑型）、numeric（数值型）、character（字符型）、factor（因子）。

函数read.table()还拥有许多微调数据导入方式的追加选项。对该函数更多的详细信息，可通过help(read.table)来获得。

3. 导入Excel文件中的数据

读取一个Excel文件中数据的最好方式，就是在Excel中将文件另存为一个用逗号分隔的文件（csv），并使用前文描述的方式将其导入R中。在Windows系统中，用户也可以使用RODBC包来访问Excel文件。电子表格文件的第一行应当包含变量/列的名称。首先，下载并安装RODBC包。执行下面的命令：

```
>install.packages("RODBC")
```

之后可以使用以下代码导入数据：

```
>library(RODBC)
>channel <- odbcConnect Excel("myfile.xls")
>mydataframe <- sqlFetch(channel, "mysheet")
>odbcclose(channel)
```

这里的myfile.xls是一个Excel文件，mysheet是要从这个工作簿中读取的工作表的名称，channel是一个由odbcConnectExcel()返回的RODBC连接对象，mydataframe是返回的数据框。RODBC也可用于从Microsoft Access导入数据。更多详情，可使用help(RODBC)来获得。Excel 2007使用了一种名为XLSX的文件格式，实质上是多个XML文件组成的压缩包。xlsx包可以用来读取这种格式的电子表格文件。在第一次使用此包之前请务必先下载并安装好。包中的函数read.xlsx()可将XLSX文件中的工作表导入为一个数据框。其最简单的调用格式是read.

xlsx(file,n)，其中file是Excel 2007工作簿的所在路径，n则为要导入的工作表序号。例如，执行以下代码：

```
>library(xlsx)
>workbook <- "c:/myworkbook.xlsx"
>mydataframe <- read.xlsx(workbook, 1)
```

该命令从位于C盘目录的工作簿myworkbook.xlsx中导入了第一个工作表，并将其保存为一个数据框mydataframe。xlsx包不仅仅可以导入数据表，它还能够创建和操作XLSX文件。

在Web数据抓取（Webscraping）的过程中，用户从互联网上提取嵌入在网页中的信息，并将其保存为可在R中做进一步的分析的数据结构。一种途径是使用函数readLines()下载网页，然后使用如gsub()和grep()这一类的函数处理它。对于结构复杂的网页，可以使用RCurl包和XML包来提取其中想要的信息。要了解更多的信息和示例，可在网站Programming with R(www.programmingr.com)上找到并阅读“Webscraping using readLines and RCurl”一文。

4. 导入SPSS数据

SPSS数据集可以通过foreign包中的函数read.spss()导入到R中，也可以使用Hmisc包中的spss.get()函数。函数spss.get()是对read.spss()的一个封装，它可以为用户自动设置后者的许多参数，让整个转换过程更加简单一致，最后得到数据分析人员所期望的结果。

首先，下载并安装Hmisc包（foreign包已被默认安装）：

```
>install.packages("Hmisc")
```

然后使用以下代码导入数据：

```
>library(Hmisc)
>mydataframe <- spss.get("mydata.sav",use.value.labels=TRUE)
```

这段代码中，mydata.sav是要导入的SPSS数据文件，use.value.labels=TRUE表示让函数将带有值标签的变量导入R中水平对应相同的因子，mydataframe是导入后的R数据框。

5. 导入SAS数据

R中设计了若干用来导入SAS数据集的函数，包括foreign包中的read.ssd()和Hmisc包中的sas.get()。但如果使用的是SAS的较新版本（SAS 9.1或更高）时，这些函数就不能正常工作了，因为R尚未跟进SAS对文件结构的改动。这里笔者推荐两种解决方案。

可以在SAS中使用PROC EXPORT将SAS数据集保存为一个逗号分隔的文本文件，然后将导出的文件读取到R中。下面是一个示例：

SAS程序：

```
proc export data=mydata
  outfile="mydata.csv"
  dbms=csv;
run;
```

R程序：

```
>mydata <- read.table("mydata.csv", header=TURE, sep=",")
```

另外，一款名为Stat/Transfer的商业软件可以完好地将SAS数据集（包括任何已知的变量格式）保存为R数据框。

6. 通过Stat/Transfer导入数据

在结束数据导入的讨论之前，值得提到一款能让上述任务的难度显著降低的商业软件。Stat/Transfer（www.stattransfer.com）是一款可在34种数据格式之间作转换的独立应用程序，其中包括R中的数据格式（见图7.4）。

此软件拥有Windows、Mac和Unix版本，并且支持目前讨论过的各种统计软件的最新版本，也可通过ODBC访问如Oracle、Sybase、Informix和DB2一类的数据库管理系统。

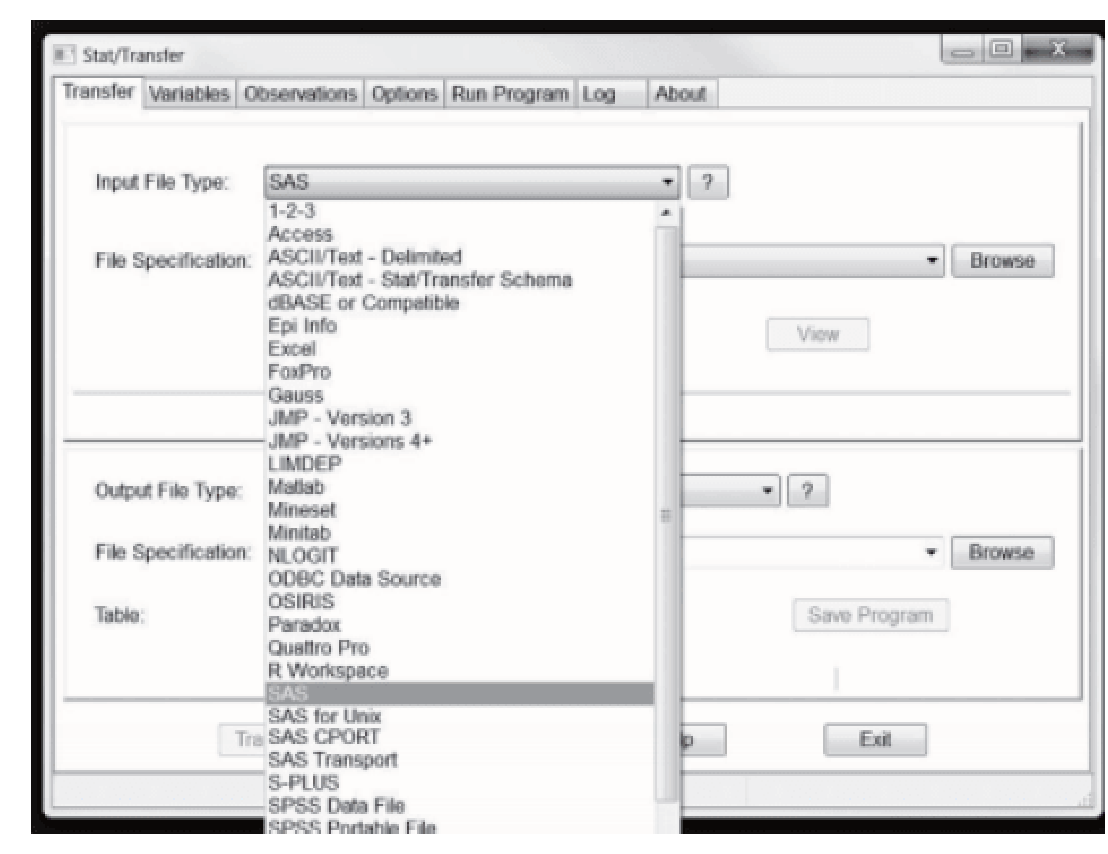


图7.4 Windows上Stat/Transfer的主对话框

7.3 绘图功能简介

R的绘图功能非常多样[①]，因为每个绘图函数都具有大量的选项使图形绘制的十分灵活多变，所以在这里不可能详细说明R在绘图方面的所有功能。若想详细了解，用户可以输入：demo(graphics)或者demo(persp)来获得详细的信息。绘图函数与本文前面描述的工作方式大为不同，绘图函数会将结果直接输出到一个“绘图设备”上，即一个绘图的窗口或是一个文件，而不是把结果赋给一个对象。绘图函数分为低级绘图函数（low-level plotting functions）和高级绘图函数（high-level plotting functions）。低级绘图函数是在现存的图形基础上添加一些元素，而高级绘图函数则是直接创建一个新的图形。此外，绘图参数（graphical parameters）可以控制绘图选项，可以用函数par修改或者使用缺省值。接下来具体介绍如何管理绘图，然后详细说明绘图函数和绘图参数及基本图形的绘制。

7.3.1 管理绘图

首先要打开多个绘图设备。绘图设备可以用适当的函数打开。如果没有打开绘图设备，绘图函数开始执行的话，R将打开一个绘图窗口来展示绘制的图形。操作系统决定了使用哪种绘图设备，在Windows系统下仍称为windows，在Unix或Linux系统下则称为x11。因为命令x11()可以作为windows()的别名，所以在Windows系统下该命令仍然有效，无论使用的是哪一种操作系统，都可以用命令x11()来打开一个绘图窗口。可以用函数postscript()、pdf()、png()等打开一个文件作为绘图设备，可以用?device命令来查看可用的绘图设备列表。最后打开的设备将成为当前的绘图设备，在该设备上显示接下来的所有的图形。如果想要显示打开的列表可

① http://www.docin.com/p-43353783.html

以用函数dev.list()。

```
> x11(); x11(); pdf()
> dev.list()
X11 X11 pdf
2 3 4
```

上面的数字是设备的编号，必须使用这些编号才能改变当前设备，若想要了解当前设备，可输入下面的命令：

```
> dev.cur()
pdf
4
```

若要改变当前的设备，可输入：

```
> dev.set(3)
X11
3
```

关闭一个设备可以用函数dev.off()。默认为关闭当前设备，否则表示关闭有自变量指定编号的设备。R然后显示新的当前设备编号。

```
> dev.off(2)
X11
3
> dev.off()
pdf
4
```

Windows版本的R，有两个特殊的功能，函数win.metafile可以打开Windows Meta-file设备，\History菜单会出现在选定绘图窗口，这个菜单中的功能可以帮助用户记录一个会话中所作的所有图形（注意：记录系统在默认状态下是关闭的，用户可以单击这个菜单下的\Recording打开它）。

若要对图形进行分割，用户可以用函数split.screen分割当前的绘图设备，例如：

```
> split.screen(c(1, 2))
```

可将设备划分为两部分，用screen(1)或者screen(2)进行选择。若要删除最后绘制的图形可以使用erase.screen()命令。而split.screen()可以作出复杂的布局，也可以划分设备的一部分。使用这些函数需要局限于图形式探索性数据分析之类的问题，和其他的函数不兼容（如layout()或coplot()），不能用于多个绘图设备。函数layout可用来将当前的图形窗口分割成多个部分，然后图形就可以一次性的显示在分割后的各部分中。它主要的自变量是一个矩阵，其中元素指示子窗口（\sub-windows"）的编号，都是整数值。例如，可以把设备划分为相等的4个部分：

```
> layout(matrix(1:4, 2, 2))
```

为更好的显现设备是如何划分的，也可以先产生这个矩阵：

```
> mat <- matrix(1:4, 2, 2)
> mat
[,1] [,2]
[1,] 1 3
[2,] 2 4
> layout(mat)
```

还可以使用函数layout.show看到创建的分割，其自变量是子窗口的个数。在图7.5所示的例子中子窗口的个数是4，执行命令：

```
> layout.show(4)
```

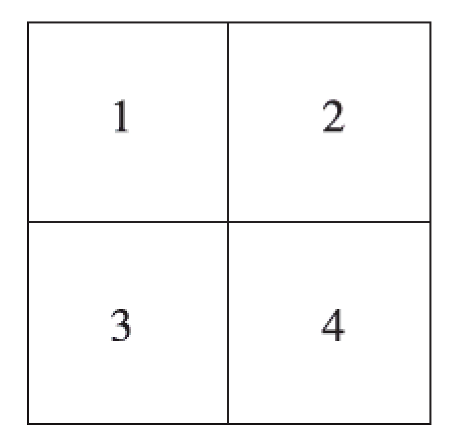


图7.5　分割后的子窗口

7.3.2　绘图函数

表7.5是R中高级绘图函数的概括[①]。

表7.5　R中高级绘图函数

函数名	说　明
pie(x)	饼图
boxplot(x)	盒形图（"box-and-whiskers"）
stripchart(x)	在一条线段上画出x的值，当样本量较小时可替代盒形图
plot(x)	以序号为横坐标，以x的元素值为纵坐标绘图
plot(x,y)	x（在x-轴上）与y（在y-轴上）的二元作图
sunflowerplot(x,y)	同上，但是将相似坐标的点作为花朵，而花瓣数目则为点的个数
matplot(x,y)	二元图，其中x的第一列对应y的第一列，x的第二列对应y的第二列，依次类推
coplot(x ~ y j z)	x与y的二元图关于z的每个数值（或数值区间）绘制
interaction.plot(f1,f2,y)	如果f1和f2是因子，作y的均值图，以f1的不同值作为x轴，而f2的不同值对应不同曲线，可以用选项fun指定y的其他的统计量（缺省计算均值，fun=mean）
fourfoldplot(x)	用四个四分之一圆显示2×2列联表情况。注意：x必须是dim=c(2,2,k)的数组，或者是dim=c(2,2)的矩阵，如果k=1
dotchart(x)	如果x是数据框，作Cleveland点图（逐行逐列累加图）
assocplot(x)	Cohen—Friendly图，显示在二维列联表中行、列变量偏离独立性的程度
termplot(mod.obj)	回归模型（mod.obj）的（偏）影响图
pairs(x)	如果x是矩阵或是数据框，作x的各列之间的二元图

① http://www.docin.com/p-43353783.html

（续表）

函数名	说　明
plot.ts(x)	如果x是类ts的对象，作x的时间序列曲线，其中x可以是多元的，但是序列的频率和时间必须相同
ts.plot(x)	同上，不过如果x是多元的，序列的频率必须相同，时间可不同
hist(x)	x的频率直方图
barplot(x)	x的值的条形图
qqnorm(x)	正态分位数，分位数图
qqplot(x,y)	y对x的分位数，分位数图
contour(x,y,z)	绘制矩阵z等高线图（在这里z表示距x–y平面的高度）x和y必须为向量，z必须为矩阵，使得dim(z)=c(length(x),length(y))，其中x和y可以省略
filled.contour(x,y,z)	同上，等高线之间的区域是彩色的，并且绘制的图例是彩色对应值的
image(x,y,z)	同上，但是可用不同色彩表示实际数据大小
persp(x,y,z)	同上，但为透视图
stars(x)	如果x是数据框或矩阵，则用星形和线段画出
mosaicplot(x)	列联表的对数线性回归残差的马赛克图
symbols(x,y,…)	由x和y给定坐标画符号，如圆，正方形，长方形，星，温度计式或者盒形图，符号的类型、大小、颜色等由另外的变量指定

在R里某些绘图函数的部分选项是一样的，用户可以在线查询每个函数的功能选项①。下面给出了一些主要的共同选项及其缺省值：

type="p"指定图形的类型。"p"：点；"l":线；"b"：点连线；"o"：同上，但是线在点上；"h"：垂直线；"S"：阶梯式，垂直线底端显示数据；"s"：同上，不过是在垂直线顶端显示数据。

xlab=,ylab=坐标轴的标签，必须是字符型值。

main=主标题，必须是字符型值。

sub=副标题，用小字体。

xlim=,ylim=指定轴的上下限，例如xlim=c(1,10)或者xlim=range(x)。

add=FALSE如果是TRUE，如果有前一个图，则叠加图形到前一个图上。

axes=TRUE如果是FALSE，不绘制轴与边框。

R里面有一套绘图函数是作用于现存的图形上的，称为低级作图命令（low–level plotting commands）。表7.6显示了一些主要的低级绘图命令。

表7.6　低级作图命令

命令	说明
points(x,y)	添加点（可以使用选项type=）
lines(x,y)	添加线，同上
segments(x0,y0,x1,y1)	从（x0,y0）各点到（x1,y1）各点画线段
arrows (x0,y0,x1,y1,angle=30, code=2)	同上，但为线添加箭头，如果code=1则在（x1,y1）处画箭头，如果code=2则在各（x0,y0）处画箭头，如果code=3则在两端都画箭头；angle用于控制箭头轴到箭头边的角度

① http://www.docin.com/p–43353783.html

（续表）

命令	说明
text(x,y,labels,…)	在（x,y）处添加用labels指定的文字，典型的用法是：plot(x,y,type="n");text(x,y,names)
mtext (text,side=3,line=0,…)	text指定的文字添加在边空，side指定添加到哪一边（参照下面的axis()）；而line指定添加的文字距离绘图区域的行数
axis(side,vect)	画坐标轴，side=1时画在下边，side=2时画在左边，side=3时画在上边，side=4时画在右边。可以选择参数at指定画刻度线的位置坐标
locator(n,type="n",…)	在用户用鼠标在图上单击n次后返回n次点击的坐标（x;y）；并可以在非缺省情况下单击处绘制符号（type="p"时）或连线（type="l"时）
abline(a,b)	绘制斜率为b和截距为a的直线
abline(v=x)	画垂直线于横坐标x处
abline(h=y)	画水平线于纵坐标y处
rect(x1,y1,x2,y2)	绘制长方形，（x1,y1）是左下角，（x2,y2）是右上角
title()	添加标题或副标题
box()	加边框在当前的图上
polygon(x,y)	绘制多边形，连接各x,y坐标确定的点
abline(lm.obj)	画回归线，由lm.obj确定的（参照第五章）
legend(x,y,legend)	在点（x,y）处添加图例，legend给定说明内容
rug(x)	用短线在x–轴上画出x数据的位置

注意，函数expression可以把自变量转换为数学公式，函数text(x,y,expression(…))可以在一个图形上加上数学公式。例如，

```
> text(x,y,expression(p==over(1,1+e^-(beta*x+alpha))))
```

在图中相应坐标点（x;y）处显示下面的方程：

$$p=\frac{1}{1+e-(\beta X+\alpha)} \qquad (7\text{–}1)$$

可以使用函数substitute和as.expression，在表达式中来代入某个变量的值，如，为了代入之前计算并储存在对象Rsquared中的R^2的值，可以：

```
> text(x,y,as.expression(substitute(R^2==r,list(r=Rsquared))))
```

则在图中相应坐标点（x;y）处会显示：

```
R²=0.9856298
```

若要求只显示3位小数，则可以将代码修改：

```
> text(x,y,as.expression(substitute(R^2==r,+ list(r=round(Rsquared,3)))))
```

则显示：R^2=0.986

7.3.3 绘图参数

除了低级作图命令之外，还可以使用绘图参数来改变图形。绘图参数[①]（不是所有参数）可以作为图形函数的选项，为了永久改变绘图参数也可以使用函数par，也就是说可以按照par指定的参数来绘制后来的图形。例如：

① http://www.docin.com/p–43353783.html

```
>par(bg="yellow")
```

上面的命令会将以后的图形的绘制背景都设置为黄色。

绘图参数有73个，其中一些功能非常相似。这些参数详细的列表可以使用命令?par来获得。

下面以一个实例详细讲解绘图命令的使用。这是一个简单的10对随机值的二维图形的例子。用以下命令生成这些值：

```
> x <- rnorm(10)
> y <- rnorm(10)
```

可以用plot()来产生所需绘制的图，命令为：

```
> plot(x,y)
```

在当前的绘图设备上将会呈现所绘制的图形，结果见图7.6。R能用“智能”的方法在默认情况下绘制图形，能自动计算坐标轴上的标记的位置，刻度摆放等，以帮助更好地理解图形。此外，绘图的方法仍然可以改变，例如，做一些个性化的调整，或满足一些刊物的要求。其中，用选项值代替缺省值来修改图形绘制是最简单的方式。由此，在上述那个例子中，可以用以下方式来改变图形：

```
>plot(x,y,xlab="Ten random values", ylab="Ten other values",
xlim=c(-2,2),ylim=c(-2,2), pch=22, col="red",
bg="yellow", bty="l", tcl=0.4,
main="How to customize a plot with R", las=1, cex=1.5)
```

结果见图7.7所示，命令中xlab和ylab默认情况下是变量的名字，用于对坐标轴的标签进行改变；xlim和ylim用于规定两个坐标轴的范围；绘图参数pch在这里用作一个选项，pch=22表示正方形，由col和bg指定其轮廓颜色和背景色；bty、tcl、las和cex的作用在图形参数表中有说明；最后，选项main用于标题的添加。此外，低级作图命令和绘图参数可以帮助用户进一步改善图形，但是需要注意的是，有一些绘图参数不允许作为plot这样的函数的自变量，此时可以通过输入多行命令，用par()来修改这些参数[①]。

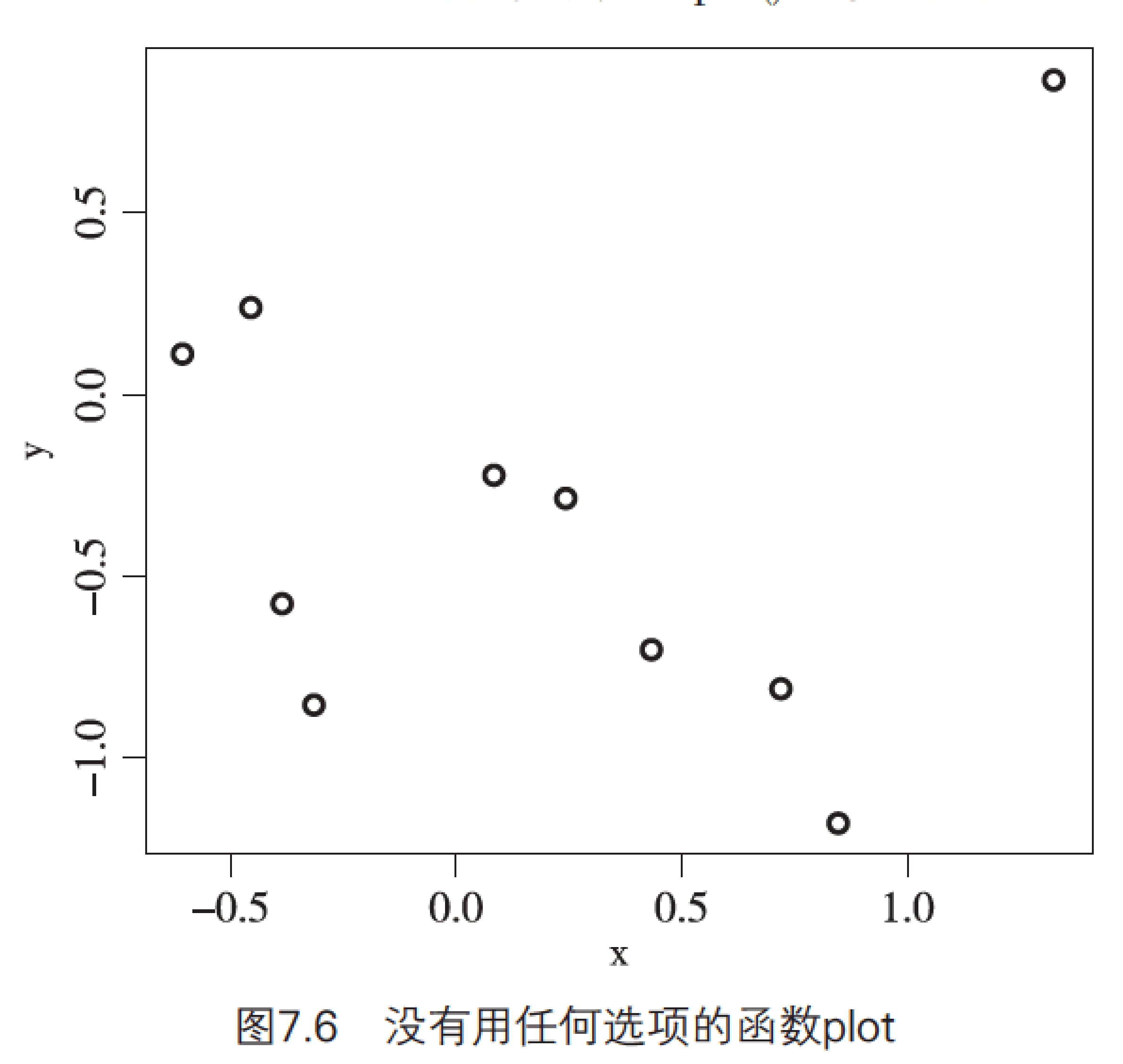


图7.6　没有用任何选项的函数plot

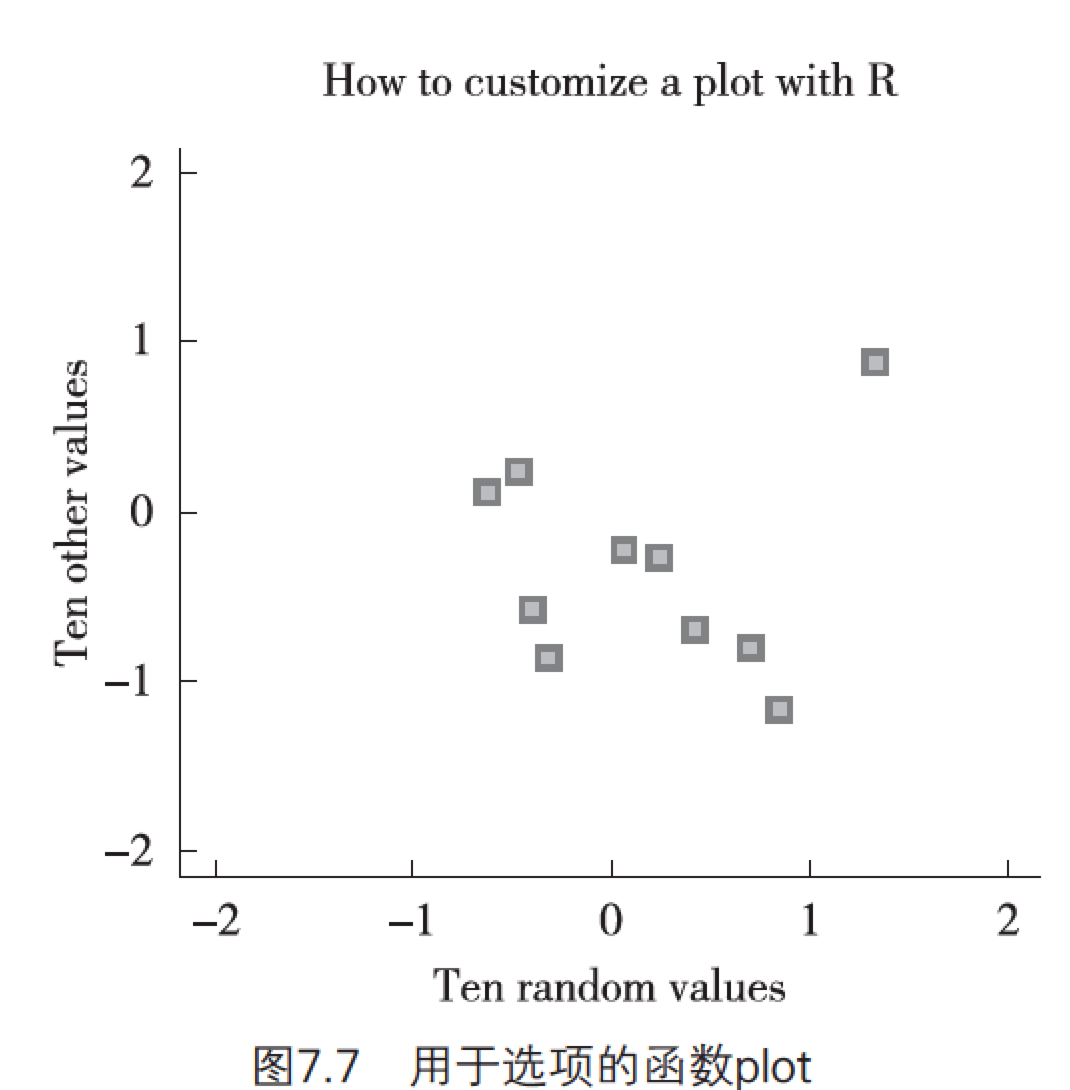


图7.7　用于选项的函数plot

① http://www.docin.com/p-43353783.html

7.3.4 基本图形

1. 条形图

条形图通过垂直的或水平的条形展示了类别型变量的分布（频数）。函数barplot()的最简单用法是：

```
barplit(height)
```

其中的height是一个向量或一个矩阵。

在接下来的示例中，将绘制一项探索类风湿性关节炎新疗法研究的结果。数据已包含在随vcd包分发的Arthritis数据框中。由于vcd包并没用包括在R的默认安装中，请确保在第一次使用之前先下载并安装它(install. packages("vcd"))。选项xlab和ylab则会分别添加x轴和y轴标签。在本例的关节炎研究中，变量Improved记录了对每位接受了安慰剂或药物治疗的病人的治疗结果。这里我们看到，28位病人有了明显改善，14人有部分改善，而42人没有改善。

```
> library(vcd)
> counts <- table(Arthritis$Improved)
> counts
```

```
None   Some Marked
  42     14     28
```

图7.8是使用垂直条形来绘制变量counts的结果，代码如下：

```
>barplot(counts, main = "Stacked Bar Plot", xlab = "Treatment",
    ylab = "Frequency", col = c("red", "yellow", "green"),
    legend = rownames(counts))
```

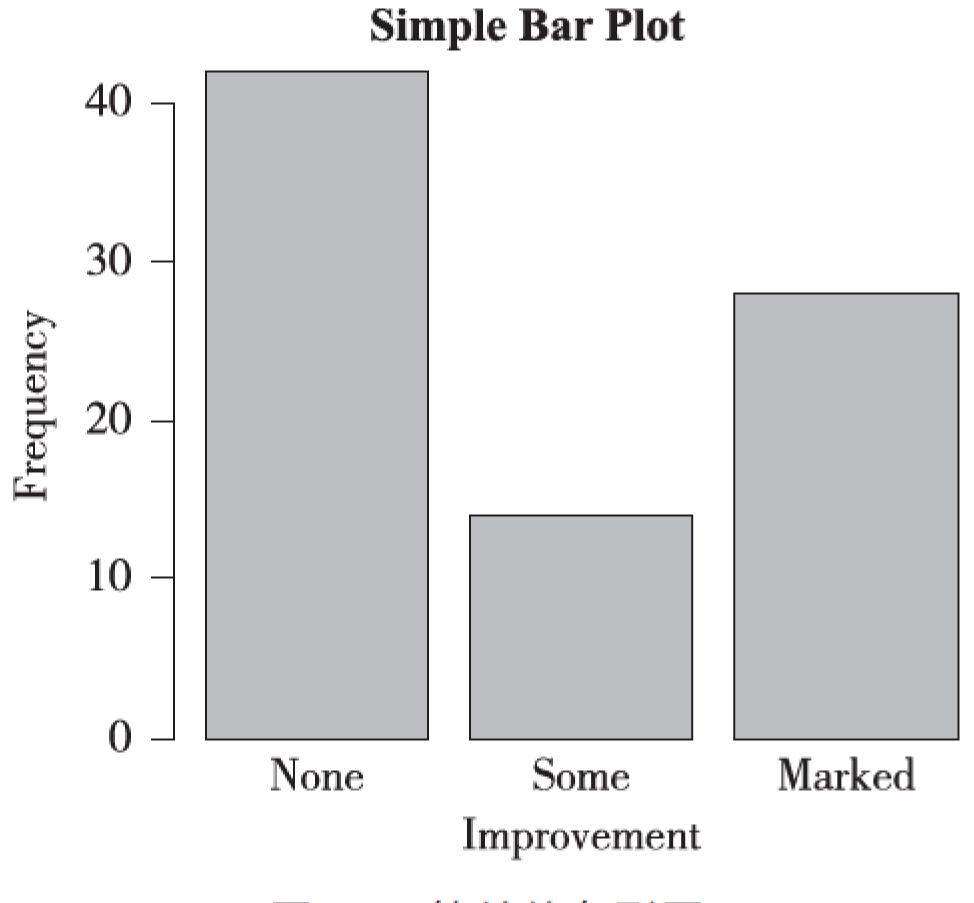


图7.8　简单的条形图

2. 直方图

直方图通过在X轴上将值域分割为一定数量的组，在Y轴上显示相应值的频数，展示连续型变量的分布。可以使用如下函数创建直方图：

```
hist(x)
```

其中的x是一个由数据值组成的数值向量。参数freq=FALSE表示根据概率密度而不是频数绘制图形。参数breaks用于控制组的数量。在定义直方图中的单元时，默认将生成等距切分。

制作图7.9所示的简单直方的代码如下：

```
> par(mfrow = c(2, 2))
> hist(mtcars$mpg)
```

图7.11所示的为添加正态密度曲线和外框的示例，生成7.10代码如下：

```
> x <- mtcars$mpg
> h <- hist(x, breaks = 12, col = "red",
    xlab = "Miles Per Gallon",
     main = "Histogram with normal curve and box")
> xfit <- seq(min(x), max(x), length = 40)
> yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
> yfit <- yfit * diff(h$mids[1:2]) * length(x)
> lines(xfit, yfit, col = "blue", lwd = 2)
> box()
```

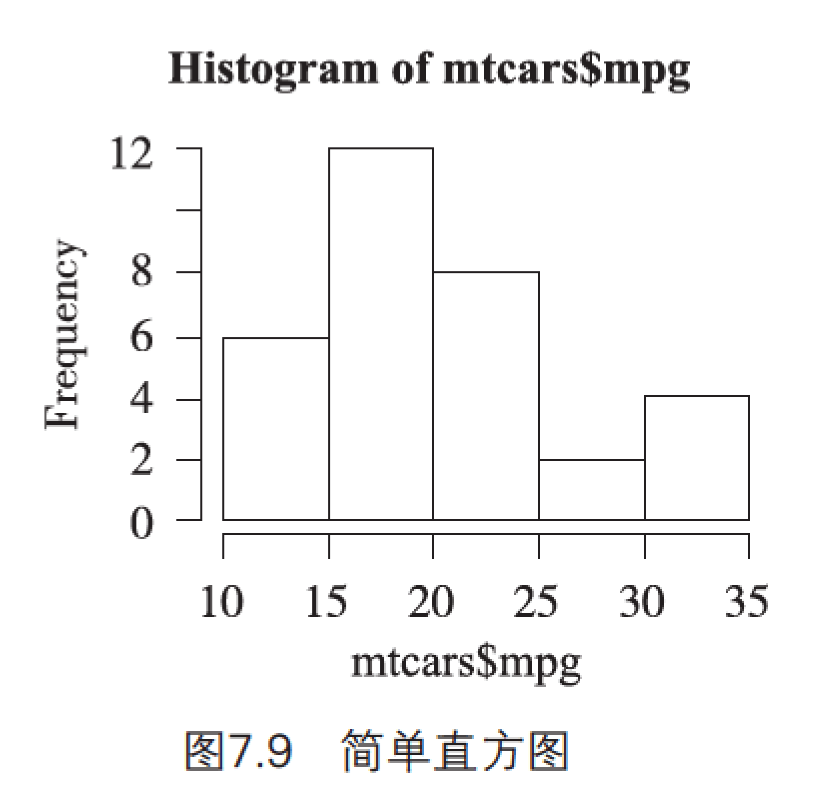


图7.9　简单直方图

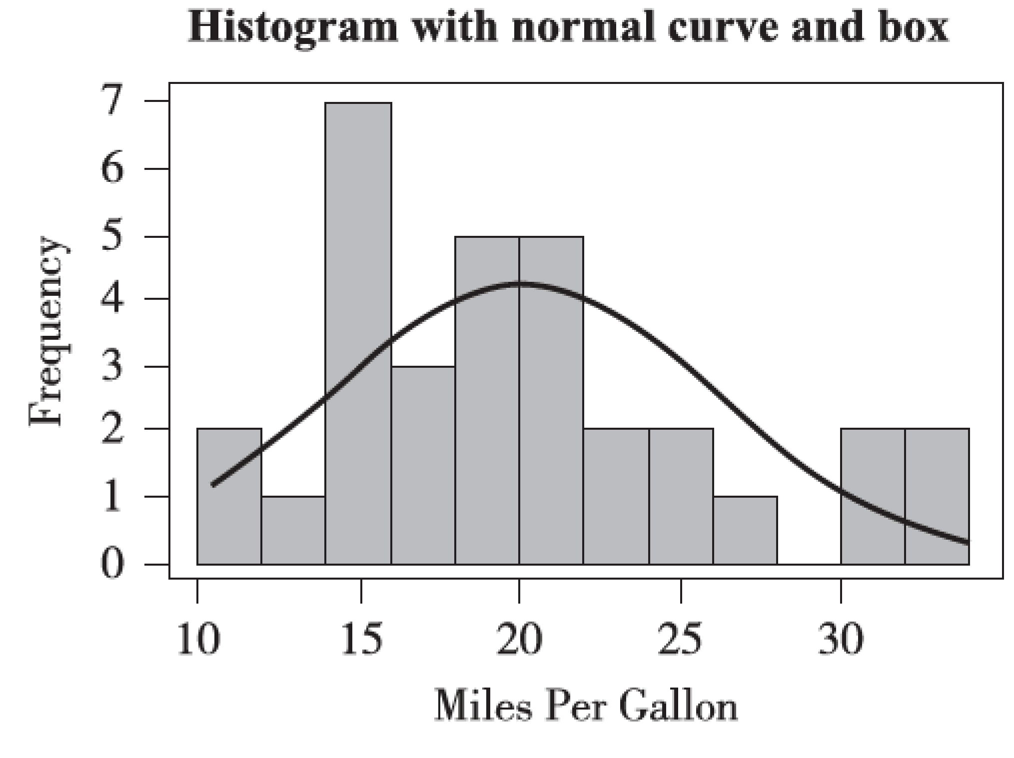


图7.10　添加正态密度曲线和外框的直方图

3. 饼图

饼图可由pie(x,labels)函数创建，其中x是一个非负数值向量，表示每个扇形的面积，而labels则表示各扇形标签的字符型向量。制作图7. 11的代码如下。

将四幅图形组合为一幅，并输入数据：

```
> par(mfrow = c(2, 2))
> slices <- c(10, 12, 4, 16, 8)
> lbls <- c("US", "UK", "Australia", "Germany", "France")
```

画一个简单的饼图：

```
> pie(slices, labels = lbls, main = "Simple Pie Chart")
```

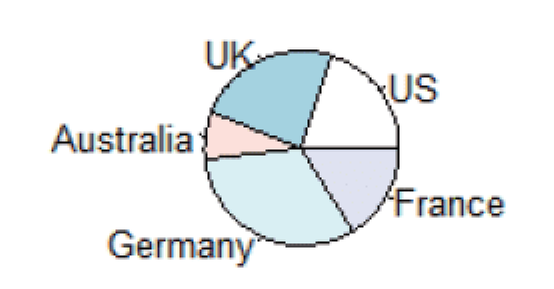
将样本数转换为比例值，并将这项信息添加到了各扇形的标签上：

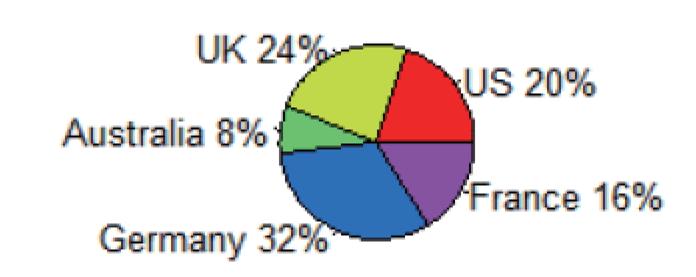
```
>pct <- round(slices/sum(slices) * 100)
```

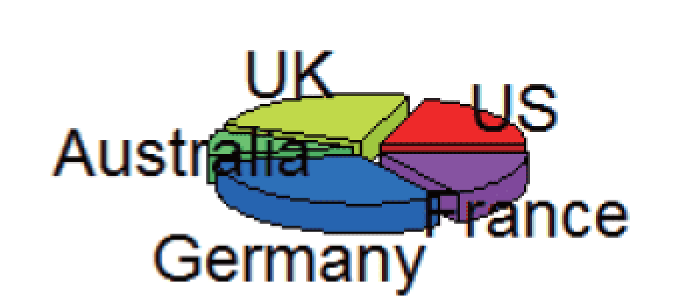
```
> lbls2 <- paste(lbls, " ", pct, "%", sep = "")
> pie(slices, labels = lbls2, col = rainbow(length(lbls)),
    main = "Pie Chart with Percentages")
```

如何从表格创建饼图：

```
> mytable <- table(state.region)
> lbls <- paste(names(mytable), "\n", mytable, sep = "")
> pie(mytable, labels = lbls,
+     main = "Pie Chart from a Table\n (with sample sizes)")
```







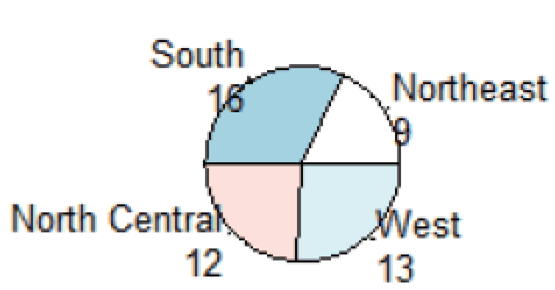


图7.11　饼图示例

4. 散点图

R中创建散点图的基础函数是plot(x, y)，其中，x和y是数值型向量，代表着图形中的（x, y）点。

加载mtcars数据框，创建了一幅基本的散点图，图形的符号是实心圆。与预期结果相同，随着车重的增加，每加仑英里数减少，虽然它们不是完美的线性关系。abline()函数用来添加最佳拟合的线性直线，而lowess()函数则用来添加一条平滑曲线。该平滑曲线拟合是一种基于局部加权多项式回归的非参数方法。制作图7.12所用的代码如下：

```
> attach(mtcars)
> plot(wt, mpg,
+      main="Basic Scatterplot of MPG vs. Weight",
+      xlab="Car Weight (lbs/1000)",
+      ylab="Miles Per Gallon ", pch=19)
> abline(lm(mpg ~ wt), col="red", lwd=2, lty=1)
> lines(lowess(wt, mpg), col="blue", lwd=2, lty=2)
```

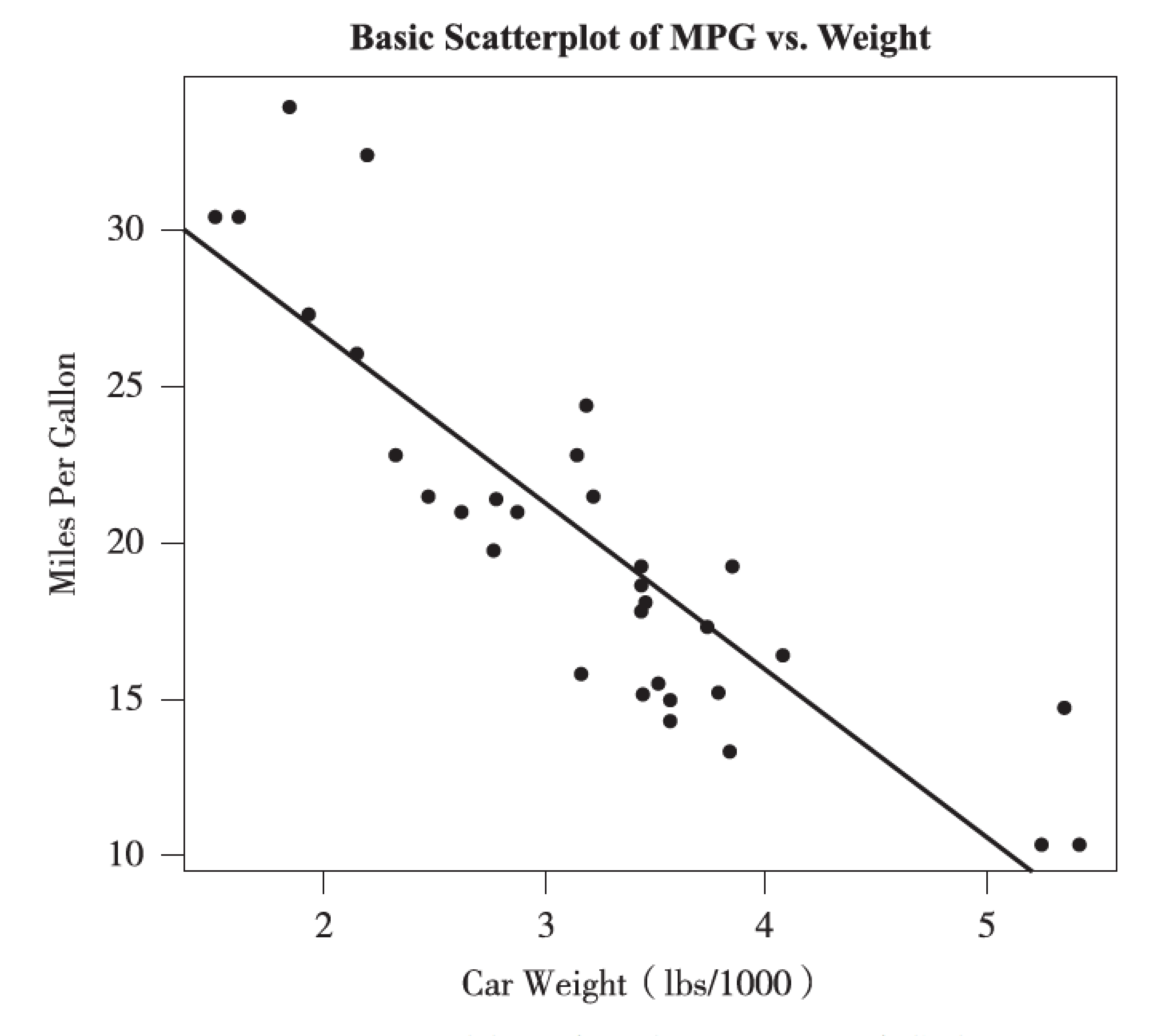


图7.12　添加了线性拟合直线和lowess拟合曲线

5. 箱线图

箱线图（又称盒形图）通过绘制连续型变量的五数情况，通过以下代码得到图7.13所示的结果：

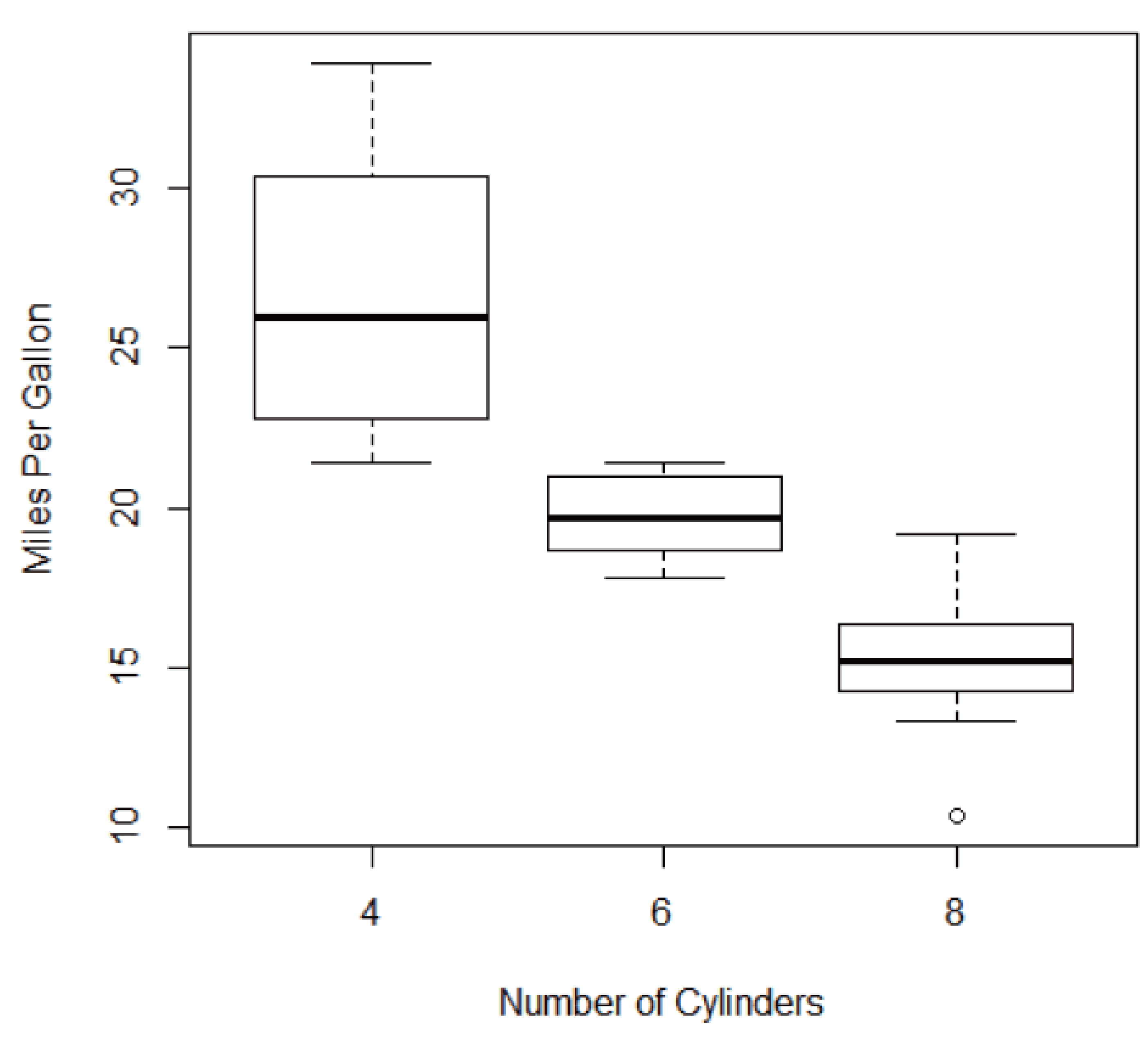


图7.13　箱线图示例

```
> boxplot(mpg ~ cyl, data = mtcars,
      main = "Car Milage Data",
      xlab = "Number of Cylinders",
ylab = "Miles Per Gallon")
```

7.4 R的初级数据分析

前面对R的绘图功能进行了介绍，但是R最厉害的是统计功能，因此本节与下一节分别从初级数据分析和高级数据分析对R的统计分析功能进行粗略的介绍。

一些基本的统计分析函数可以从包stats中获得，例如，包括方差分析、广义线性模型和最小二乘法回归的线性模型、非线性最小二乘法、多元分析、经典的假设检验、汇总统计、时间序列分析、层次聚类和统计分布。上述统计方法以外的统计方法还可以从其他R包中获得。与基本R安装同时发布的统计包称为推荐包，称其他包为捐献包并且用户需要自己安装。下面介绍在所有统计分析中非常有用的两个概念：公式（formulae）和泛型函数（genericfunctions）[①]。

1. 公式

因为几乎所有函数的符号都一样（基本上趋同，也有例外），所以公式在R统计分析里非常重要，是关键元素。Y ~ model是公式的典型形式，其中model要为其中一些项估计参数，是一些元素项的集合，这些元素项被由有特殊涵义的运算符连接，而y是响应变量。

a:b	a和b的交互效应
a+b	a和b的相加效应
a*b	相加和交互效应（等价于a+b+a:b）
–b	去掉因子b的影响，如:（a+b+c）^2–a:b等价于a+b+c+a:c+b:c
1	y ~ 1拟合一个没有因子影响的模型（仅仅是截距）
–1	y ~ x–1表示通过原点的线性回归（等价于y ~ x+0或者0+y ~ x）
^n	包含所有的直到n阶的交互作用，即（a+b+c）^2等价于a+b+c+a:b+a:c+b:c
poly(a,n)	a的n价多项式
X	如X是一个矩阵，这将反映各列的相加效应，即X[,1]+X[,2]+…+X[,ncol(X)]；还可以通过索引向量选择特定列进行分析（如X[,2:4]）
b%in%a	b和a的嵌套分类设计（等价于a+a:b，或者a/b）
offset(…)	在向模型中增加一个影响因子但不估计任何参数（如，offset(3*x)）

我们可以看出，在R公式里面采用的运算符和表达式里面使用的运算符含义不尽相同。例如，公式$y \sim x_1+x_2$表示模型$y=\beta_1x_1+\beta_2x_2+\alpha$，而不是（如果+采用它常规的含义）$y=\beta（x_1+x_2）+\alpha$。我们可以使用函数I:公式$y \sim \mathrm{I}(x_1+x_2)$表示模型$y=\beta(x_1+x_2)+\alpha$以便可以在公式中使用常规的运算符。相似的，我们可以使用公式$y \sim \mathrm{poly}(x,2)$（而非$y \sim x+x^2$）来定义模型$y=\beta_1x+\beta_2x^2+\alpha$。但

① http://blog.sina.com.cn/s/blog_7dd658650100tfoq.html

是，我们也可以在公式中包含一些函数以便对变量进行一定的转换。对于方差分析，aov()定义随机效应时用了一个特别的语法规则。例如，y ~ a+Error(b)可以对固定项a和随机项b的相加效应进行表示。

2. 泛型函数

泛型（generic）[①]就是用来解析结果的，对特定的类对象有特定的行为的函数。R函数将输入对象的属性作为输入参数，这一点不同于很多其他的统计编程语言。其中最应该关注的一个属性则是类。R统计函数通常返回的对象的类名与函数名相同，例如，aov返回类“aov”的对象，lm返回类“lm”的对象。

泛型函数的优势在于一个函数对所有类的使用格式都是一样的。例如，summary是最常用的用于解析统计分析结果的R函数，它可以显示较为细致的结果。无论作为参数的对象是“lm”类（线性模型）还是“aov”类（方差分析），显示的信息是不一样的。

R还有一个重要性质是，一个包含分析结果的对象常常是一个列表对象，它的类定义决定了它的结果展示方式，即输入参数的对象类型决定一个函数的行为。如表7.7所示列出了一些泛型函数，主要用于提取分析结果对象的信息：

```
> mod <- lm(y ~ x)
> df.residual(mod)
[1] 8
```

表7.7　提取分析结果对象的信息的主要泛型函数

函数名	说　明
print	返回简单的汇总信息
summary	返回较为详细的汇总信息
residuals	返回残差
df.residual	返回残差的自由度
fitted	返回拟合值
deviance	返回方差
coef	返回被估计的系数以及标准差
AIC	计算Akaike信息准则（Akaike information criterion，AIC）（依赖于logLik()）
logLik	计算返回参数数目和对数似然值

泛型函数通常是调用自变量所属类的对应函数，其中调用的函数称为方法（method），很少对对象进行操作。简单来说，一个方法的构建方式是generic.cls，其中cls是对象的类。例如，以summary为例，下面是对应的方法:

```
> apropos("^summary")
[1] "summary" "summary.aov"
[3] "summary.aovlist" "summary.connection"
[5] "summary.data.frame" "summary.default"
[7] "summary.factor" "summary.glm"
```

① http://blog.sina.com.cn/s/blog_7dd65865010100s1.html

```
[9] "summary.glm.null" "summary.infl"
[11] "summary.lm" "summary.lm.null"
[13] "summary.manova" "summary.matrix"
[15] "summary.mlm" "summary.packageStatus"
[17] "summary.POSIXct" "summary.POSIXlt"
[19] "summary.table"
```

在线性回归和方差分析中泛型函数的行为是不同的，通过下面的例子，可以看出：

```
> x <- y <- rnorm(5);
> lm.spray <- lm(y ~ x)
> names(lm.spray)
[1] "coefficients" "residuals" "effects"
[4] "rank" "fitted.values" "assign"
[7] "qr" "df.residual" "xlevels"
[10] "call" "terms" "model"
> names(summary(lm.spray))
[1] "call" "terms" "residuals"
[4] "coefficients" "sigma" "df"
[7] "r.squared" "adj.r.squared" "fstatistic"
[10] "cov.unscaled"
```

表7.8给出了其他一些泛型函数，可以对分析结果对象做一些补充分析，其主要参数一般是其分析结果对象，但有些情况下需要一些额外的参数，如泛型函数predict或update。

表7.8　泛型函数的额外参数

函数名	说　明
predict	借助拟合的模型计算一个新的数据集的预测值
update	使用新的数据或者公式拟合一个模型
anova	计算一个或多个模型的方差/残差分析表
drop1	连续测试所有可以从模型中移除的元素项
add1	连续测试所有可以加入模型的元素项
step	通过AIC（调用add1和drop1）选择一个模型

除此之外，还有很多效用函数和图形函数，效用函数可用于从模型对象或公式中提取信息，如可用来查找一个特定公式拟合的线性模型中的线性依赖项的函数alias。图形函数，如可以显示多种多样的诊断图的plot，还有上面例子中的termplot，尽管后面一个函数不是泛型函数但它调用了泛型函数predict。

7.4.1　描述性统计分析

1. 单组数据的描述性统计

描述性统计中，样本的观测值中含有总体各方面的信息，它来自总体，信息较为分散，

显得杂乱无章。为了能够反映总体的各项特征，需要将这些分散在样本中的有关总体的信息集中起来，对样本进行加工得到统计量[①]。在描述性统计量的计算方面，R中有很多的选择。下面先从基础安装中包含的函数入手学习，然后再学习用户贡献包中的扩展函数。

本节将使用Motor Trend杂志的车辆路试（mtcars）数据集。这里的关注焦点是每加仑汽油行驶英里数（mpg）、马力（hp）和车重（wt）。

```
> vars <- c("mpg", "hp", "wt")
> head(mtcars[vars])
```

```
                   mpg  hp   w
Mazda RX4         21.0 110 2.62
Mazda RX4 Wag     21.0 110 2.87
Datsun 710        22.8  93 2.32
Hornet 4 Drive    21.4 110 3.21
Hornet Sportabout 18.7 175 3.44
```

首先查看所有32种车型的描述性统计量，然后按照变速箱类型（am）和汽缸数（cyl）考察描述性统计量。变速箱类型是一个以0表示自动挡、1表示手动挡来编码的二分变量，而汽缸数可为4、5或6。

对于基础安装，可以使用summary()函数来获取描述性统计量。

```
> summary(mtcars[vars])
```

```
      mpg             hp              wt
 Min.   :10.40   Min.   : 52.0   Min.   :1.51
 1st Qu.:15.43   1st Qu.: 96.5   1st Qu.:2.58
 Median :19.20   Median :123.0   Median :3.32
 Mean   :20.09   Mean   :146.7   Mean   :3.21
 3rd Qu.:22.80   3rd Qu.:180.0   3rd Qu.:3.61
```

summary()函数提供了最小值、最大值、四分位数和数值型变量的均值，以及因子向量和逻辑型向量的频数统计。可以使用apply()函数或sapply()函数计算所选择的任意描述性统计量。对于sapply()函数，其使用格式为：

```
>Sapply(x,FUN,options)
```

其中的x是用户定义的数据框（或矩阵），FUN为一个任意的函数。如果指定了options，它们将被传递给FUN。用户定义可以在这里插入的典型函数有mean、sd、var、min、max、median、length、range和quantile。函数fivenum()可返回图基五数概括（Tukey’s five-number summary，即最小值、下四分位数、中位数、上四分位数和最大值）。

令人惊讶的是，基础安装并没有提供偏度和峰度的计算函数，不过用户可以自行添加如下代码：

```
> mystats <- function(x, na.omit = FALSE) {
+     if (na.omit)
+         x <- x[!is.na(x)]
+     m <- mean(x)
+     n <- length(x)
+     s <- sd(x)
```

① http://www.btdcw.com/btd-3421104f767f5acfa1c7cd3d-2.html.

```
+     skew <- sum((x - m)^3/s^3)/n
+     kurt <- sum((x - m)^4/s^4)/n - 3
+     return(c(n = n, mean = m, stdev = s, skew = skew, kurtosis = kurt))
+ }
> sapply(mtcars[vars], mystats)
```

```
               mpg          hp          wt
n        32.000000  32.0000000 32.00000000
mean     20.090625 146.6875000  3.21725000
stdev     6.026948  68.5628685  0.97845744
skew      0.610655   0.7260237  0.42314646
kurtosis -0.372766  -0.1355511 -0.02271075
```

对于样本中的车型，每加仑汽油行驶英里数的平均值为20.1，标准差为6.0。分布呈现右偏（偏度+0.61），并且较正态分布稍平（峰度~0.37）。如果对数据进行绘图操作，那这些特征就更显而易见了。注意，如果只希望单纯地忽略缺失值，那么应当使用：

```
sapply (mtcars[vars], mystats, na.omit=TRUE)
```

2. 分组计算的描述性统计量

在比较多组个体或观测时，关注的焦点经常是各组的描述性统计信息，而不是样本整体的描述性统计信息。同样地，在R中完成这个任务有很多种方法。我们将以获取变速箱类型各水平的描述性统计量开始。可以使用aggregate()函数来分组获取描述性统计量，代码如下：

```
> aggregate(mtcars[vars], by = list(am = mtcars$am), mean)
```

```
  am      mpg       hp       wt
1  0 17.14737 160.2632 3.768895
2  1 24.39231 126.8462 2.411000
```

```
> aggregate(mtcars[vars], by = list(am = mtcars$am), sd)
```

```
  am      mpg       hp        wt
1  0 3.833966 53.90820 0.7774001
2  1 6.166504 84.06232 0.6169816
```

注意list(am=mtcars$am)的使用。如果使用的是list(mtcars$am)，则am列将被标注为Group.1而不是am。使用这个赋值指定了一个更有帮助的列标签。

psych包中的describe.by()函数可计算和describe相同的描述性统计量，只是按照一个或多个分组变量分层，使用psych包中的describe.by()分组计算概述统计量

```
> describe.by(mtcars[vars], mtcars$am)
```

```
group: 0
    vars  n   mean    sd median trimmed   mad   min    max  range  s
mpg    1 19  17.15  3.83  17.30   17.12  3.11 10.40  24.40  14.00  0
hp     2 19 160.26 53.91 175.00  161.06 77.10 62.00 245.00 183.00 -0
wt     3 19   3.77  0.78   3.52    3.75  0.45  2.46   5.42   2.96  0
    kurtosis    se
mpg    -0.80  0.88
hp     -1.21 12.37
wt      0.14  0.18
---------------------------------------------------------
group: 1
    vars  n   mean    sd median trimmed   mad   min    max  range sk
mpg    1 13  24.39  6.17  22.80   24.38  6.67 15.00  33.90  18.90 0.
hp     2 13 126.85 84.06 109.00  114.73 63.75 52.00 335.00 283.00 1.
wt     3 13   2.41  0.62   2.32    2.39  0.68  1.51   3.57   2.06 0.
    kurtosis    se
mpg    -1.46  1.71
hp      0.56 23.31
wt     -1.17  0.17
```

与前面的示例不同，describe.by()函数不允许指定任意函数，所以它的普适性较低。若存在一个以上的分组变量，可以使用list(groupvar1, groupvar2,..., groupvarN)来表示它们。但这仅在分组变量交叉后不出现空白单元时有效。

数据分析人员对于展示哪些描述性统计量以及结果采用什么格式都有着自己的偏好，这也许就是有如此多不同方法的原因。用户可以选择最适合的方式，或是创造属于自己的方法。

7.4.2　频数表和列联表

在本节中，我们将着眼于类别型变量的频数表和列联表，以及相应的独立性检验、相关性的度量、图形化展示结果的方法。除了使用基础安装中的函数外，还将使用vcd包和gmodels包中的函数。下面的示例中，假设A、B和C代表类别型变量。

本节中的数据来自vcd包中的Arthritis数据集。这份数据来自Kock&Edward(1988)，表示了一项风湿性关节炎新疗法的双盲临床实验的结果。前几个观测是这样的：

```
> library(vcd)
> head(Arthritis)
```

```
  ID Treatment  Sex Age Improved
1 57   Treated Male  27     Some
2 46   Treated Male  29     None
3 77   Treated Male  30     None
4 17   Treated Male  32   Marked
5 36   Treated Male  46   Marked
6 23   Treated Male  58   Marked
```

治疗情况（安慰剂治疗、用药治疗）、性别（男性、女性）和改善情况（无改善、一定程度的改善、显著改善）均为类别型因子。

R中提供了用于创建频数表和列联表的若干种方法。其中最重要的函数已列于表7.9中。

表7.9　创建频数表和列联表方法

函　数	描　述
table(var1, var2, ⋯, varN)	使用 N 个类别型变量（因子）创建一个 N 维列联表
xtabs(formula, data)	根据一个公式和一个矩阵或数据框创建一个 N 维列联表
prop.table(table, margins)	依margins定义的边际列表将表中条目表示为分数形式
margin.table(table, margins)	依margins定义的边际列表计算表中条目的和
addmargins(table,margins)	将概述边margins（默认是求和结果）放入表中
ftable(table)	创建一个紧凑的“平铺”式列联表

接下来，将逐个使用以上函数来探索类别型变量。首先考察简单的频率表，接下来是二维列联表，最后是多维列联表。第一步是使用table()或xtabs()函数创建一个表，然后使用其他函数处理它。

一维列联表可以使用table()函数生成简单的频数统计表。示例如下：

```
> mytable <- with(Arthritis, table(Improved))
> mytable
```

```
Improved
  None   Some Marked
    42     14     28
```

可以用prop.table()将这些频数转化为比例值：

```
> prop.table(mytable)
```

```
Improved
     None      Some    Marked
0.5000000 0.1666667 0.3333333
```

或使用prop.table()*100转化为百分比：

```
> prop.table(mytable)*100
```

```
Improved
    None     Some   Marked
50.00000 16.66667 33.33333
```

这里可以看到，有50%的研究参与者获得了一定程度或者显著的改善（16.7 + 33.3）。对于二维列联表，table()函数的使用格式为：

```
mytable<-table(A,B)
```

其中的A是行变量，B是列变量。除此之外，xtabs()函数还可使用公式风格的输入创建列联表，格式为：

```
mytable<-xtabs(~A+B, datamydata)
```

其中的mydata是一个矩阵或数据框。总的来说，要进行交叉分类的变量应出现在公式的右侧（即 ~ 符号的右方），以+作为分隔符。若某个变量写在公式的左侧，则其为一个频数向量（在数据已经被表格化时很有用）。

对于Arthritis数据，有：

```
> mytable <- xtabs(~ Treatment+Improved, data=Arthritis)
> mytable
```

对于Arthritis数据，有：

```
> mytable <- xtabs(~ Treatment+Improved, data=Arthritis)
> mytable
```


```
         Improved
Treatment None Some Marked
  Placebo   29    7      7
  Treated   13    7     21
```

可以使用margin.table()和prop.table()函数分别生成边际频数和比例。行和与行比例可以这样计算：

```
> margin.table(mytable, 1)
```

```
Treatment
Placebo Treated
     43      41
```

```
> prop.table(mytable, 1)
```

```
         Improved
Treatment      None      Some    Marked
  Placebo 0.6744186 0.1627907 0.1627907
  Treated 0.3170732 0.1707317 0.5121951
```

下标1指代table()语句中的第一个变量。观察表格可以发现，与接受安慰剂的个体中有显著改善的16%相比，接受治疗的个体中的51%的个体病情有了显著的改善。

列和与列比例可以这样计算：

```
> margin.table(mytable, 2)
```

```
Improved
  None   Some Marked
    42     14     28
```

```
> prop.table(mytable, 2)
```

```
          Improved
Treatment      None      Some    Marked
  Placebo 0.6904762 0.5000000 0.2500000
  Treated 0.3095238 0.5000000 0.7500000
```

这里的下标2指代table()语句中的第二个变量，如果有两个以上的类别型变量，那么就是在处理多维列联表，table()和xtabs()都可以基于三个或更多的类别型变量生成多维列联表。margin.table()、prop.table()和addmargins()函数可以自然地推广到高于二维的情况。另外，ftable()函数可以以一种紧凑而吸引人的方式输出多维列联表。

以下代码生产了三维分组各单元格的频数，同时演示了如何使用ftable()函数输出更加紧凑和吸引人的表格。

```
> mytable <- xtabs(~ Treatment+Sex+Improved, data=Arthritis)
```

```
, , Improved = None

         Sex
Treatment Female Male
  Placebo     19   10
  Treated      6    7

, , Improved = Some

         Sex
Treatment Female Male
  Placebo      7    0
  Treated      5    2

, , Improved = Marked

         Sex
Treatment Female Male
  Placebo      6    1
  Treated     16    5
```

```
>ftable( mytable)
```

```
                  Improved None Some Marked
Treatment Sex
Placebo   Female             19    7      6
          Male               10    0      1
Treated   Female              6    5     16
          Male                7    2      5
```

以下代码为治疗情况（Treatment）、性别（Sex）和改善情况（Improved）生成了边际频数。使用公式 ~ Treatement+Sex+Improve创建了这个表，所以Treatment需要通过下标1来引用；Sex通过下标2来引用；Improve通过下标3来引用。

```
> margin.table(mytable, 1)
```

```
Treatment
Placebo Treated
     43      41
```

```
> margin.table(mytable, 2)
```

```
Sex
Female    Male
    59      25
```

```
> margin.table(mytable, 3)
```

```
Improved
  None   Some Marked
    42     14     28
```

下面的代码为治疗情况（Treatment）× 改善情况（Improved）分组的边际频数，由不同性别（Sex）的单元汇总而成。

```
> margin.table(mytable, c(1,3))
```

```
         Improved
Treatment None Some Marked
  Placebo   29    7      7
  Treated   13    7     21
```

每个Treatment × Sex组合中改善情况为None、Some和Marked患者的比例。在这里可以看到治疗组的男性中有36%有了显著改善，女性为59%。总而言之，比例将被添加到不在prop.table()调用中的下标上。

```
> ftable(prop.table(mytable, c(1, 2)))
```

```
                  Improved       None       Some      Mar
Treatment Sex
Placebo   Female            0.59375000 0.21875000 0.18750
          Male              0.90909091 0.00000000 0.09090
Treated   Female            0.22222222 0.18518519 0.59259
          Male              0.50000000 0.14285714 0.35714
```

```
> ftable(addmargins(prop.table(mytable,c(1,2)),3))
```

```
                  Improved       None       Some     Marked
Treatment Sex
Placebo   Female            0.59375000 0.21875000 0.18750000 1.00000
          Male              0.90909091 0.00000000 0.09090909 1.00000
Treated   Female            0.22222222 0.18518519 0.59259259 1.00000
          Male              0.50000000 0.14285714 0.35714286 1.00000
```

列联表可以看出组成表格的各种变量组合的频数或比例，可能还会对列联表中的变量是否相关或独立感兴趣。可使用卡方独立性检验chisq.text()和Fisher精确检验fisher.text()等进行独立性的检验。

7.4.3　相关分析

根据变量间的相互关系可分为两种类型[①②]，相关关系和函数关系。相关关系是相关分析的研究对象，指两个变量的数值变化存在的依存关系不完全确定，它们之间的数值不能用方

① http://wenku.baidu.com/view/187565115f0e7cd18425366f.html

② http://blog.sina.com.cn/s/blog_92dbc65401015lo6.html

程表示出来，但可用某种相关性度量来刻划；函数关系是回归分析的研究对象，指变量之间存在的相互依存关系，它们之间的关系可以用某一方程*y=f(x)*表达出来。相关的种类繁多，按照不同的标准有不同的划分。根据不同的相关形式，可以划分为线性相关和非线性相关；根据不同的相关方向，可以划分为正相关和负相关；根据不同的相关程度，可以划分为不相关、不完全相关和完全相关；根据涉及变量的多少，可分为一元相关和多元相关；根据影响因素的不同，可分为单相关和复相关。

R可以计算多种相关系数，包括Pearson相关系数、Spearman相关系数、Kendall相关系数、偏相关系数、多分格（polychoric）相关系数和多系列（polyserial）相关系数。下面就来介绍这些相关系数。

1. Pearson、Spearman和Kendall相关

Pearson积差相关系数衡量了两个定量变量之间的线性相关程度。Spearman等级相关系数则衡量分级定序变量之间的相关程度。Kendall's Tau相关系数也是一种非参数的等级相关度量。cor()函数可以计算这三种相关系数，而cov()函数可用来计算协方差。两个函数的参数有很多，其中与相关系数的计算有关的参数可以简化为：

```
cor(x,use= ,method= )
```

x：矩阵或数据；use：指定缺失数据的处理方式。可选的方式为all.obs（假设不存在缺失数据，遇到缺失数据时将报错）、everything（遇到缺失数据时，相关系数的计算结果将被设为missing）、complete.obs（行删除）以及pairwise.complete.obs（成对删除，pairwisedeletion）；method：指定相关系数的类型。可选类型为pearson、spearman或kendall默认参数为use="everything"和method="pearson"。

```
> states <- state.x77[, 1:6]
> cov(states)
```

```
             Population      Income   Illiteracy     Life Exp      Murder
Population 19931683.7588 571229.7796  292.8679592 -407.8424612 5663.523714
Income       571229.7796 377573.3061 -163.7020408  280.6631837 -521.894286
Illiteracy      292.8680   -163.7020    0.3715306   -0.4815122    1.581776
Life Exp       -407.8425    280.6632   -0.4815122    1.8020204   -3.869480
Murder         5663.5237   -521.8943    1.5817755   -3.8694804   13.627465
HS Grad       -3551.5096   3076.7690   -3.2354694    6.3126849  -14.549616
               HS Grad
Population -3551.509551
Income      3076.768980
Illiteracy    -3.235469
Life Exp       6.312685
Murder       -14.549616
HS Grad       65.237894
```

```
> cor(states)
```

```
            Population     Income Illiteracy    Life Exp     Mur
Population  1.00000000  0.2082276  0.1076224 -0.06805195  0.3436
Income      0.20822756  1.0000000 -0.4370752  0.34025534 -0.2300
Illiteracy  0.10762237 -0.4370752  1.0000000 -0.58847793  0.7029
Life Exp   -0.06805195  0.3402553 -0.5884779  1.00000000 -0.7808
Murder      0.34364275 -0.2300776  0.7029752 -0.78084575  1.0000
HS Grad    -0.09848975  0.6199323 -0.6571886  0.58221620 -0.4879
               HS Grad
Population -0.09848975
Income      0.61993232
Illiteracy -0.65718861
Life Exp    0.58221620
Murder     -0.48797102
HS Grad     1.00000000
```

```
> cor(states, method="spearman")
```

```
            Population     Income Illiteracy   Life Exp     Murder    HS Grad
Population   1.0000000  0.1246098  0.3130496 -0.1040171  0.3457401 -0.3833649
Income       0.1246098  1.0000000 -0.3145948  0.3241050 -0.2174623  0.5104809
Illiteracy   0.3130496 -0.3145948  1.0000000 -0.5553735  0.6723592 -0.6545396
Life Exp    -0.1040171  0.3241050 -0.5553735  1.0000000 -0.7802406  0.5239410
Murder       0.3457401 -0.2174623  0.6723592 -0.7802406  1.0000000 -0.4367330
HS Grad     -0.3833649  0.5104809 -0.6545396  0.5239410 -0.4367330  1.0000000
```

第一个语句是计算方差和协方差，第二个语句则计算Pearson积差相关系数，第三个语句计算Spearman等级相关系数。在该例中，可以看到收入和高中毕业率之间存在很强的正相关，而文盲率和预期寿命之间存在很强的负相关。

2. 偏相关

偏相关是指在控制一个或多个定量变量时，另外两个定量变量之间的相互关系。可以使用ggm包中的pcor()函数计算偏相关系数。ggm包没有被默认安装，在第一次使用之前需要先进行安装。函数调用格式为：

```
pcor(u,s)
```

其中的u是一个数值向量，前两个数值表示要计算相关系数的变量下标，其余的数值为条件变量（即要排除影响的变量）的下标。S为变量的协方差阵。这个示例有助于阐明用法：

```
> library(ggm)
> #在控制了收入、文盲率和高中毕业率时
> #人口和谋杀率的偏相关系数
> pcor(c(1, 5, 2, 3, 6), cov(states))
[1] 0.3462724
```

本例中，在控制了收入、文盲率和高中毕业率的影响时，人口和谋杀率之间的相关系数为0.346。偏相关系数常用于社会科学的研究中。

3. 其他类型的相关

polycor包中的hetcor()函数可以计算一种混合的相关矩阵，其中包括数值型变量的Pearson积差相关系数、数值型变量和有序变量之间的多系列相关系数、有序变量之间的多分格相关系数以及二分变量之间的四分相关系数。多系列、多分格和四分相关系数都假设有序变量或二分变量由潜在的正态分布导出。请参考此程序包所附文档以了解更多内容。

4. 相关性的显著性检验

在计算好相关系数以后，如何对它们进行统计显著性检验呢，常用的原假设为变量间不相关（即总体的相关系数为0）。可使用cor.test()函数对单个的Pearson、Spearman和Kendall相关系数进行检验。简化后的使用格式为：

```
cor.test(x,y,alternative=  ,method=  )
```

其中的x和y为要检验相关性的变量，alternative则用来指定进行双侧检验或单侧检验（取值为"two.side"、"less"或"greater"），而method用以指定要计算的相关类型（"pearson"、"kendall"或"spearman"）。当研究的假设为总体的相关系数小于0时，请使用alternative="less"。在研究的假设为总体的相关系数大于0时，应使用alternative="greater"。在

默认情况下，假设为alternative="two.side"（总体相关系数不等于0）。

```
> cor.test(states[, 3], states[, 5])

        Pearson's product-moment correlation

data:  states[, 3] and states[, 5]
t = 6.8479, df = 48, p-value = 1.258e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5279280 0.8207295
sample estimates:
      cor
0.7029752
```

以上代码检验的原假设是：预期寿命和谋杀率的Pearson相关系数为0。假设总体的相关度为0，则预计在一千万次中只会有少于一次的机会见到0.703这样大的样本相关度（即p=1.258e-8）。由于这种情况几乎不可能发生，故拒绝原假设，从而支持了要研究的假设，即预期寿命和谋杀率之间的总体相关度不为0。

7.4.4 *t*检验

在研究中最常见的行为就是对两个组进行比较。接受某种新药治疗的患者是否较使用某种现有药物的患者表现出了更大程度的改善？某种制造工艺是否较另外一种工艺制造出的不合格品更少？两种教学方法中哪一种更有效？如果你的结果变量是类别型的，那么可以直接使用7.3节中阐述的方法。这里我们将关注结果变量为连续型的组间比较，并假设其呈正态分布。为了阐明方法，将使用MASS包中的UScrime数据集。它包含了1960年美国47个州的刑罚制度对犯罪率影响的信息。我们感兴趣的结果变量为Prob（监禁的概率）、U1（14～24岁年龄段城市男性失业率）和U2（35～39岁年龄段城市男性失业率）。类别型变量So（指示该州是否位于南方的指示变量）将作为分组变量使用。数据的尺度已被原始作者缩放过。

1. 独立样本的*t*检验

如果一个人在美国的南方犯罪，是否更有可能被判监禁？这里比较的对象是美国南方和非南方各州，因变量为监禁的概率。一个针对两组的独立样本*t*检验可以用于检验两个总体的均值相等的假设。这里假设两组数据是独立的，并且是从正态总体中抽得。检验的调用格式为：

```
t.test(y~x,data)
```

其中的y是一个数值型变量，x是一个二分变量。调用格式为：

```
t.test(y1,y2)
```

其中的y1和y2为数值型向量（即各组的结果变量）。可选参数data的取值为一个包含了这些变量的矩阵或数据框。与其他多数统计软件不同的是，这里的*t*检验默认假定方差不相等，并使用Welsh的修正自由度。可以添加一个参数var.equal=TRUE以假定方差相等，并使用合并方差估计。默认的备择假设是双侧的（即均值不相等，但大小的方向不确定）。可以添加一个参数alternative="less"或alternative="greater"来进行有方向的检验。

在下列代码中，我们使用了一个假设方差不等的双侧检验，比较了南方（group1）和非南方（group0）各州的监禁概率：

```
> t.test(Prob ~ So, data=UScrime)
```

```
        Welch Two Sample t-test

data:  Prob by So
t = -3.8954, df = 24.925, p-value = 0.0006506
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03852569 -0.01187439
sample estimates:
mean in group 0 mean in group 1
     0.03851265      0.06371269
```

可以拒绝南方各州和非南方各州拥有相同监禁概率的假设（$p < .001$）。

2. 非独立样本的*t*检验

再举个例子，可能会问：较年轻（14 ~ 24岁）男性的失业率是否比年长（35 ~ 39岁）男性的失业率更高？在这种情况下，这两组数据并不独立。不能说亚拉巴马州的年轻男性和年长男性的失业率之间没有关系。在两组的观测之间相关时，获得的是一个非独立组设计（dependent groups design）。前—后测设计（pre-post design）或重复测量设计（repeated measures design）同样也会产生非独立的组。

非独立样本的*t*检验假定组间的差异呈正态分布。对于本例，检验的调用格式为：

```
t.test(y1,y2,paired=TRUE)
```

其中的y1和y2为两个非独立组的数值向量。结果如下：

```
> library(MASS)
> sapply(UScrime[c("U1", "U2")], function(x) (c(mean = mean(x), sd =
sd(x))))
```

```
           U1       U2
mean 95.46809 33.97872
sd   18.02878  8.44545
```

```
> with(UScrime, t.test(U1, U2, paired = TRUE))
```

```
        Paired t-test

data:  U1 and U2
t = 32.4066, df = 46, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 57.67003 65.30870
sample estimates:
mean of the differences
               61.48936
```

差异的均值（61.5）足够大，可以保证拒绝年长和年轻男性的平均失业率相同的假设。年轻男性的失业率更高。事实上，若总体均值相等，获取一个差异如此大的样本的概率小于0.000 000 000 000 000 22。

7.4.5 回归分析

与回归分析相比，相关分析不能回答在两个变量之间存在相关关系时，它们之间是如何联系的，即无法找出刻画它们之间因果关系的函数关系，而只能得出两个变量之间是否存在

相关关系。回归分析就可以解决这一问题，先从一元线性回归讲起。

1. 简单线性回归

设变量*x*和*y*之间存在一定的相关关系，回归分析方法即找出*Y*的值是如何随*X*的值的变化而变化的规律，称*Y*为因变量（或响应变量），*X*为自变量（或解释变量）。

最简单的拟合回归模型函数是lm()，格式为：

```
myfit<-lm(formula,data)
```

其中，formula指要拟合的模型形式，data是一个数据框，包含了用于拟合模型的数据。结果对象（本例中是myfit）存储在一个列表中，包含了所拟合模型的大量信息。表达式（formula）形式如下：

$$Y \sim X_1+X_2+\cdots+X_k$$

~左边为响应变量，右边为各个预测变量，预测变量之间用+号分隔。除了lm()，表7.10还列出了其他一些对做简单或多元回归分析有用的函数。拟合模型后，将这些函数应用于lm()返回的对象，可以得到更多额外的模型信息。

表7.10　对拟合线性模型非常有用的其他函数

函　数	用　途
summary()	展示拟合模型的详细结果
coefficients()	列出拟合模型的模型参数（截距项和斜率）
confint()	提供模型参数的置信区间（默认95%）
fitted()	列出拟合模型的预测值
residuals()	列出拟合模型的残差值
anova()	生成一个拟合模型的方差分析表，或者比较两个或更多拟合模型的方差分析表
vcov()	列出模型参数的协方差矩阵
AIC()	输出赤池信息统计量
plot()	生成评价拟合模型的诊断图
predict()	用拟合模型对新的数据集预测响应变量值

当回归模型包含一个因变量和一个自变量时，称为简单线性回归。当只有一个预测变量，但同时包含变量的幂（比如，X、X_2、X_3）时，称之为多项式回归。当有不止一个预测变量时，则称为多元线性回归。现在，首先从一个简单的线性回归例子开始介绍。

例：bschool是美国60个著名商学院的数据，包括的变量有GMAT分数、学费、进入MBA前后的工资等，现在仅研究进入MBA前后的工资变化。执行下述命令，可得如图7.14所示的结果。

```
y=read.csv("bschool0.txt")
attach(y)
plot(SalaPreMBA,SalaPostMBA,xlab="SalaryPreMBA",ylab="SalaryPostMBA")
```

可以看出，进入MBA前工资高的，毕业后也高。

```
> y=read.csv("bschool.txt",header=T,sep=",")
> lm=lm(SalaPostMBA~SalaPreMBA,y)
> summary(lm)
```

```
Call:
lm(formula = SalaPostMBA ~ SalaPreMBA, data = y)

Residuals:
    Min      1Q  Median      3Q     Max
-32.877  -5.952  -0.087   6.802  23.636

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.4026     6.8394  -1.667    0.101
SalaPreMBA    2.8290     0.1535  18.434   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.1 on 58 degrees of freedom
Multiple R-squared:  0.8542,    Adjusted R-squared:  0.8517
F-statistic: 339.8 on 1 and 58 DF,  p-value: < 2.2e-16
```

执行下述命令得到如图7.15所示的图表。

```
> s<-y$SalaPreMBA
> t<-y$SalaPostMBA
> plot(s,t,xlab="Salary Pre MBA",ylab="Salary Post MBA")
> abline(lm,col="blue",lty=1)
```

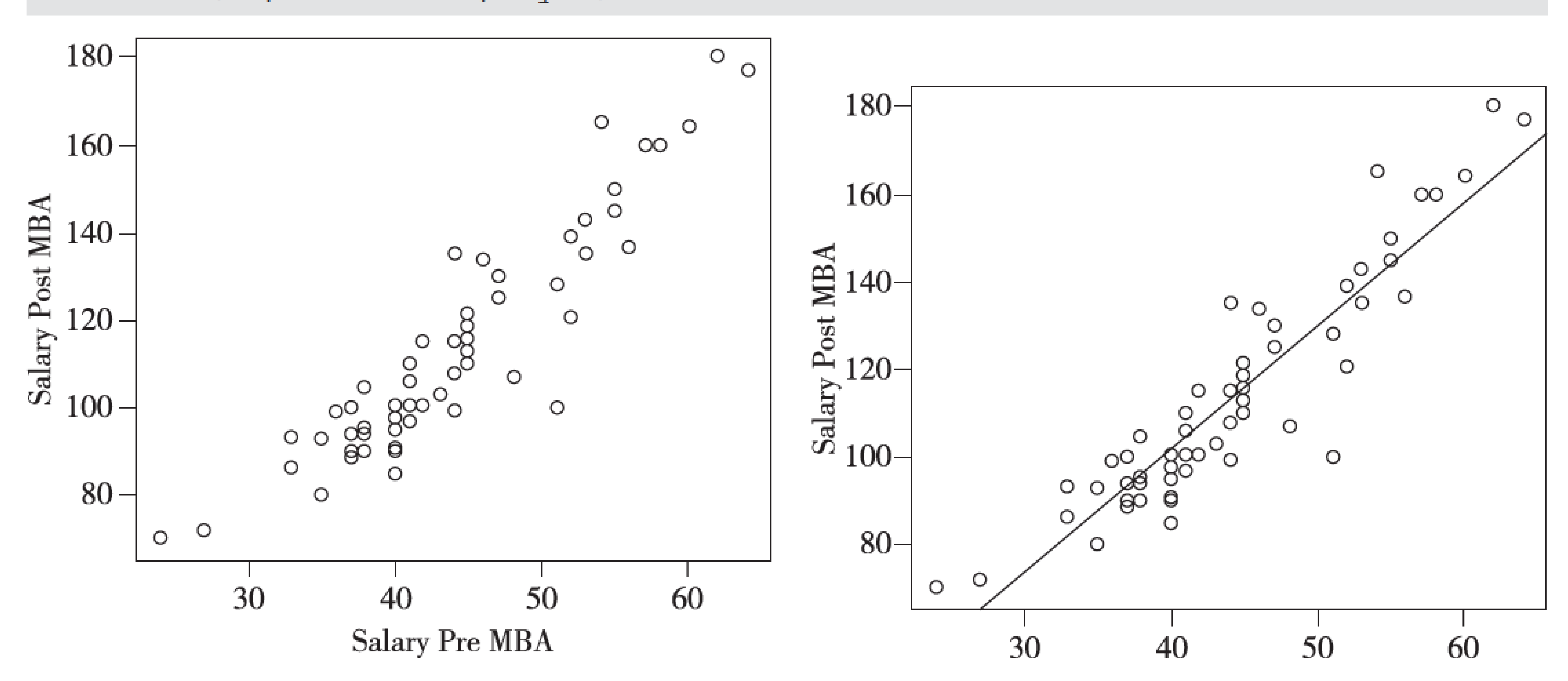


图7.14　进入MBA前后的工资变化　　图7.15　进入MBA前的工资和得到MBA后的工资回归直线

根据计算，找到进入MBA前的工资和得到MBA之后的工资的回归直线。R软件的输出给出来的截距-11.4026和斜率2.829，该直线的方程为：

$$Y=-11.40+2.83x$$

2. 多元线性回归

当预测变量不止一个时，简单线性回归就变成了多元线性回归，分析也稍微复杂些。从技术上来说，多项式回归可以算是多元线性回归的特例，二次回归有两个预测变量（X和X_2），三次回归有三个预测变量（X、X_2和X_3）。现在让我们看一个更一般的例子。

以基础包中的state.x77数据集为例，在本例中想探究一个州的犯罪率和其他因素的关系，包括人口、文盲率、平均收入和结霜天数（温度在冰点以下的平均天数）。

因为lm()函数需要一个数据框（state.x77数据集是矩阵），为了以后处理方便，需要做如下转化：

```
states <- as.data.frame(state.x77[, c("Murder", "Population", "Illiteracy",
"Income", "Frost")])
```

这行代码创建了一个名为states的数据框，包含了我们感兴趣的变量。本章的余下部分，我们都将使用这个新的数据框。多元回归分析中，第一步最好检查一下变量间的相关性。cor()函数提供了二变量之间的相关系数，car包中scatterplotMatrix()函数则会生成散点图矩阵图7.16所示。

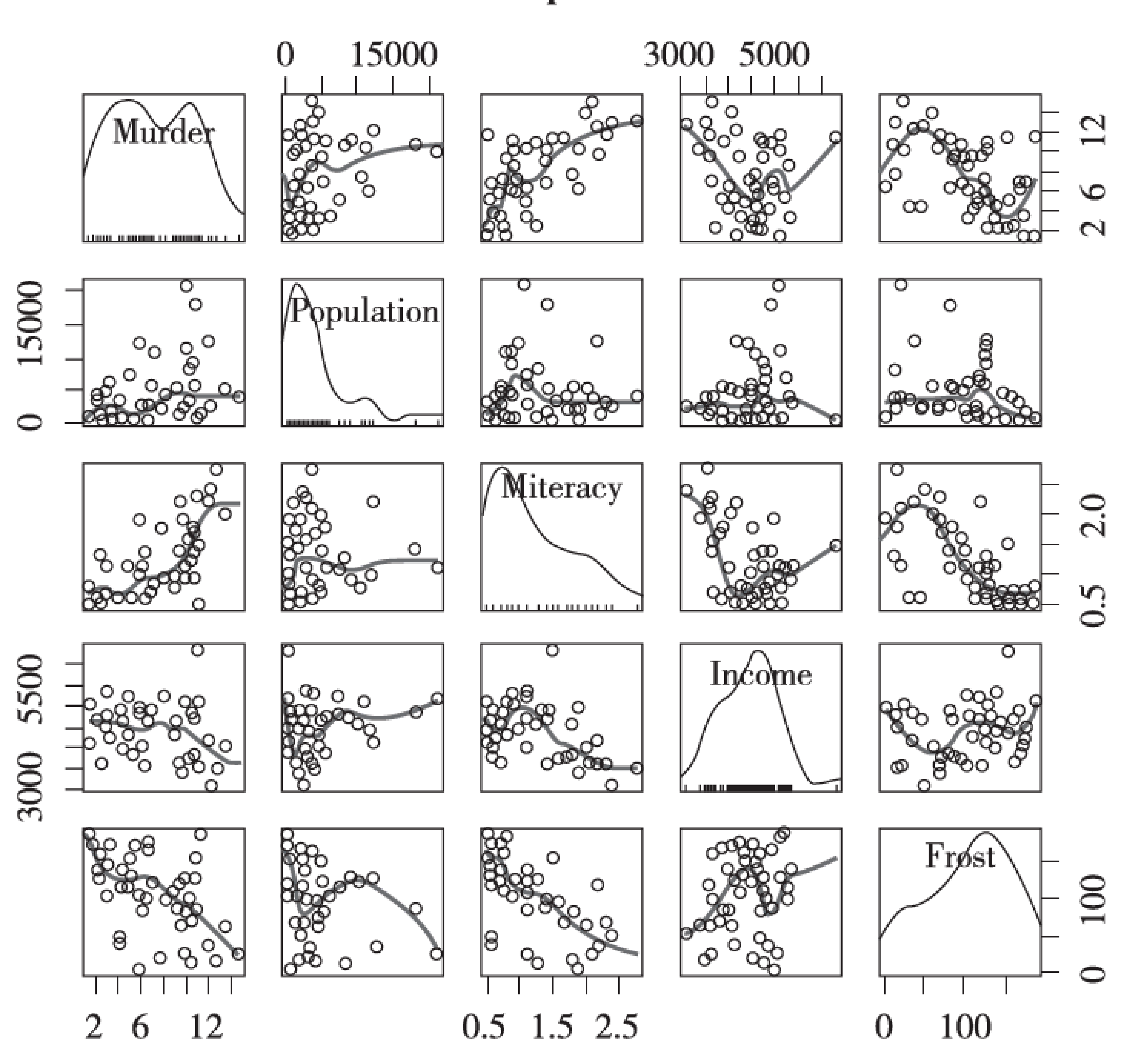


图7.16　州府数据中因变量与自变量的散点图矩阵

检测二变量关系：

```
> cor(states)
```

```
              Murder Population Illiteracy     Income      Frost
Murder      1.0000000  0.3436428  0.7029752 -0.2300776 -0.5388834
Population  0.3436428  1.0000000  0.1076224  0.2082276 -0.3321525
Illiteracy  0.7029752  0.1076224  1.0000000 -0.4370752 -0.6719470
Income     -0.2300776  0.2082276 -0.4370752  1.0000000  0.2262822
Frost      -0.5388834 -0.3321525 -0.6719470  0.2262822  1.0000000
```

```
>library(car) > scatterplotMatr
ix(states, spread = FALSE, lty.smooth = 2, main = "Scatterplot Matrix")
```

scatterplotMatrix()函数默认在非对角线区域绘制变量间的散点图，并添加平滑（loess）

和线性拟合曲线。对角线区域绘制每个变量的密度图和轴须图。从图中可以看到，谋杀率是双峰的曲线，每个预测变量都一定程度上出现了偏斜。谋杀率随着人口和文盲率的增加而增加，随着收入水平和结霜天数增加而下降。同时可看出，越冷的州府文盲率越低，收入水平越高。

现在使用lm()函数拟合多元线性回归模型：

```
> fit <- lm(Murder ~ Population + Illiteracy + Income + Frost, data =
states)
> summary(fit)
```

```
Call:
lm(formula = Murder ~ Population + Illiteracy + Income + Frost,
    data = states)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7960 -1.6495 -0.0811  1.4815  7.6210

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.235e+00  3.866e+00   0.319   0.7510
Population  2.237e-04  9.052e-05   2.471   0.0173 *
Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
Income      6.442e-05  6.837e-04   0.094   0.9253
Frost       5.813e-04  1.005e-02   0.058   0.9541
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.535 on 45 degrees of freedom
Multiple R-squared:  0.567,     Adjusted R-squared:  0.5285
F-statistic: 14.73 on 4 and 45 DF,  p-value: 9.133e-08
```

当预测变量不止一个时，回归系数的含义为：一个预测变量增加一个单位，其他预测变量保持不变时，因变量将要增加的数量。例如本例中，文盲率的回归系数为4.14，表示控制人口、收入和温度不变时，文盲率上升1%，谋杀率将会上升4.14%，它的系数在$p<0.001$的水平下显著不为0。相反，Frost的系数没有显著不为0（$p=0.954$），表明当控制其他变量不变时，Frost与Murder不呈线性相关。总体来看，所有的预测变量解释了各州谋杀率57%的方差。

以上分析中，没有考虑预测变量的交互项，在接下来的章节中，我们将考虑一个包含此因素的例子。

3. 回归诊断

使用lm()函数来拟合回归模型，通过summary()等函数获取模型的参数和相关统计量。但是，没有任何输出能告诉你模型是否合适，对模型参数推断的信心依赖于它在多大程度上满足OLS模型统计假设。为数据的无规律性或者错误设定了预测变量与响应变量的关系，都将使你的模型产生巨大的偏差。一方面，可能得出某个预测变量与响应变量无关的结论，但事实上，它们相关；另一方面，情况可能恰好相反。当模型应用到真实世界中时，预测效果可能很差，误差显著。

回归诊断技术向用户提供了评价回归模型适用性的必要工具，它能帮助发现并纠正问题。这里的探讨使用了R基础包中的函数的标准方法。

R基础安装中提供了大量检验回归分析中统计假设的方法。最常见的方法就是对lm()函数返回的对象使用plot()函数，可以生成评价模型拟合情况的四幅图形。虽然这些标准的诊断图

形很有用，但是R中还有更好的工具可用，例如car包提供了大量函数，大大增强了拟合和评价回归模型的能力，在此就不详细阐述。

（1）正态性

与基础包中的plot()函数相比，qqplot()函数提供了更为精确的正态假设检验方法，它画出了在$n-p-1$个自由度的t分布下的学生化残差（studentized residual，也称学生化删残差或折叠化残差）图形，其中n是样本大小，p是回归参数的数目（包括截距项）。

（2）误差的独立性

判断因变量值（或残差）是否相互独立，最好的方法是依据收集数据方式的先验知识。例如，时间序列数据通常呈现自相关性——相隔时间越近的观测相关性大于相隔越远的观测。car包提供了一个可做Durbin–Watson检验的函数，能够检测误差的序列相关性。

（3）线性

通过成分残差图（component plus residual plot）也称偏残差图（partial residual plot），可以看看因变量与自变量之间是否呈非线性关系，也可以看看是否有不同于已设定线性模型的系统偏差，图形可用car包中的crPlots()函数绘制。

（4）同方差性

car包提供了两个有用的函数，可以判断误差方差是否恒定。ncvTest()函数生成一个计分检验，零假设为误差方差不变，备择假设为误差方差随着拟合值水平的变化而变化。若检验显著，则说明存在异方差性（误差方差不恒定）。

（5）多重共线性

多重共线性（multicollinearity），它会导致模型参数的置信区间过大，使单个系数解释起来很困难。多重共线性可用统计量VIF（Variance Inflation Factor，方差膨胀因子）进行检测。VIF的平方根表示变量回归参数的置信区间能膨胀为与模型无关的预测变量的程度（因此而得名）。car包中的vif()函数可以提供VIF值。一般原则下，vif>2就表明存在多重共线性问题。

4. 模型的选择

当选定一个模型，并且用数据拟合时，并不一定所有的变量都显著，或者说并不一定所有的系数都有意义，这时需要一边回归一边检验，进行所谓逐步回归，这个方法或者从只有常数项开始，逐步把显著的变量加入；或者从包含所有变量的模型考试，逐步把不显著的变量减去；也可以同时有加有减的双向逐步回归。

逐步回归法的实现依据增删变量的准则不同而不同。MASS包中的stepAIC()函数可以实现逐步回归模型（向前、向后和向前向后），依据的是精确AIC准则。下面的例子中，采用的是向后回归。

```
> library(MASS)
> fit1 <- lm(Murder ~ Population + Illiteracy + Income + Frost, data =
states)
> stepAIC(fit, direction = "backward")
```

```
Start:  AIC=97.75
Murder ~ Population + Illiteracy + Income + +Frost

             Df Sum of Sq    RSS     AIC
- Frost       1     0.021 289.19  95.753
- Income      1     0.057 289.22  95.759
<none>                    289.17  97.749
- Population  1    39.238 328.41 102.111
- Illiteracy  1   144.264 433.43 115.986

Step:  AIC=95.75
Murder ~ Population + Illiteracy + Income

             Df Sum of Sq    RSS     AIC
- Income      1     0.057 289.25  93.763
<none>                    289.19  95.753
- Population  1    43.658 332.85 100.783
- Illiteracy  1   236.196 525.38 123.605

Step:  AIC=93.76
Murder ~ Population + Illiteracy

             Df Sum of Sq    RSS     AIC
<none>                    289.25  93.763
- Population  1    48.517 337.76  99.516
- Illiteracy  1   299.646 588.89 127.311
```

开始时模型包含4个（全部）预测变量，然后在每一步中，AIC列提供了删除一个行中变量后模型的AIC值，<none>中的AIC值表示没有变量被删除时模型的AIC。第一步，Frost被删除，AIC从97.75降低到95.75；第二步，删除Income，AIC继续下降到为93.76；然后再删除变量将会增加AIC，因此终止选择过程。逐步回归法其实存在争议，虽然它可能会找到一个好的模型，但是不能保证模型就是最佳模型，因为不是每一个可能的模型都被评价了。为克服这个限制，便有了全子集回归法。

7.4.6　方差分析

方差分析又称“*F*检验”“变异数分析”[①]，可对两个或两个以上的样本均数的差别进行显著性检验。它是一种分析各个自变量对因变量的影响的方法，其自变量是定性变量的因子及可能出现的称为协变量的定量变量。首先自变量的取值不同，因变量的值也会变化，方差分析可对其进行分解，从而得出每一个自变量对结果都有一份贡献。其次把剩下的不能用已知原因解释的当做随机误差。然后对各自变量和随机误差的贡献进行*F*检验，从而输出*F*–值和检验的一些*p*–值，来判断该自变量的不同水平对因变量的变化是否有显著贡献。最后会得出一个方差分析表来表示分析结果。

1. 单因素方差分析

单因素方差分析，比较分类因子定义的两个或多个组别中的因变量均值。以multcomp包中的cholesterol数据集为例（取自Westfall，Tobia，Rom，Hochberg，1999），50个患者均接受降低胆固醇药物治疗（trt）五种疗法中的一种疗法。其中三种治疗条件使用药物相同，分别是20mg一天一次（1time）、10mg一天两次（2times）和5mg一天四次（4times）。剩下的两种方式（drugD和drugE）代表候选药物。哪种药物疗法降低胆固醇（响应变量）最多？

```
> library(multcomp)
> attach(cholesterol)
> table(trt)
```

① http://wenku.baidu.com/view/38c4f9640b1c59eef8c7b4f5.html

```
trt
 1time 2times 4times  drugD  drugE
    10     10     10     10     10
```

```
> aggregate(response, by = list(trt), FUN = mean)
```

```
  Group.1        x
1   1time  5.78197
2  2times  9.22497
3  4times 12.37478
4   drugD 15.36117
5   drugE 20.94752
```

```
> aggregate(response, by = list(trt), FUN = sd)
```

```
  Group.1        x
1   1time 2.878113
2  2times 3.483054
3  4times 2.923119
4   drugD 3.454636
5   drugE 3.345003
```

```
> fit <- aov(response ~ trt)
> summary(fit)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
trt          4 1351.4   337.8   32.43 9.82e-13 ***
Residuals   45  468.8    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> library(gplots)
>plotmeans(response ~ trt, xlab = "Treatment", ylab = "Response", main =
"Mean Plot\nwith 95% CI")
> detach(cholesterol)
```

从输出结果可以看到，每10个患者接受其中一个药物疗法。均值显示drugE降低胆固醇最多，而1time降低胆固醇最少，各组的标准差相对恒定，在2.88到3.48间浮动。ANOVA对治疗方式（trt）的F检验非常显著（$p<0.0001$），说明五种疗法的效果不同。gplots包中的plotmeans()可以用来绘制带有置信区间的组均值图形，如图7.17所示，图形展示了带有95%的置信区间的各疗法均值，可以清楚看到它们之间的差异。

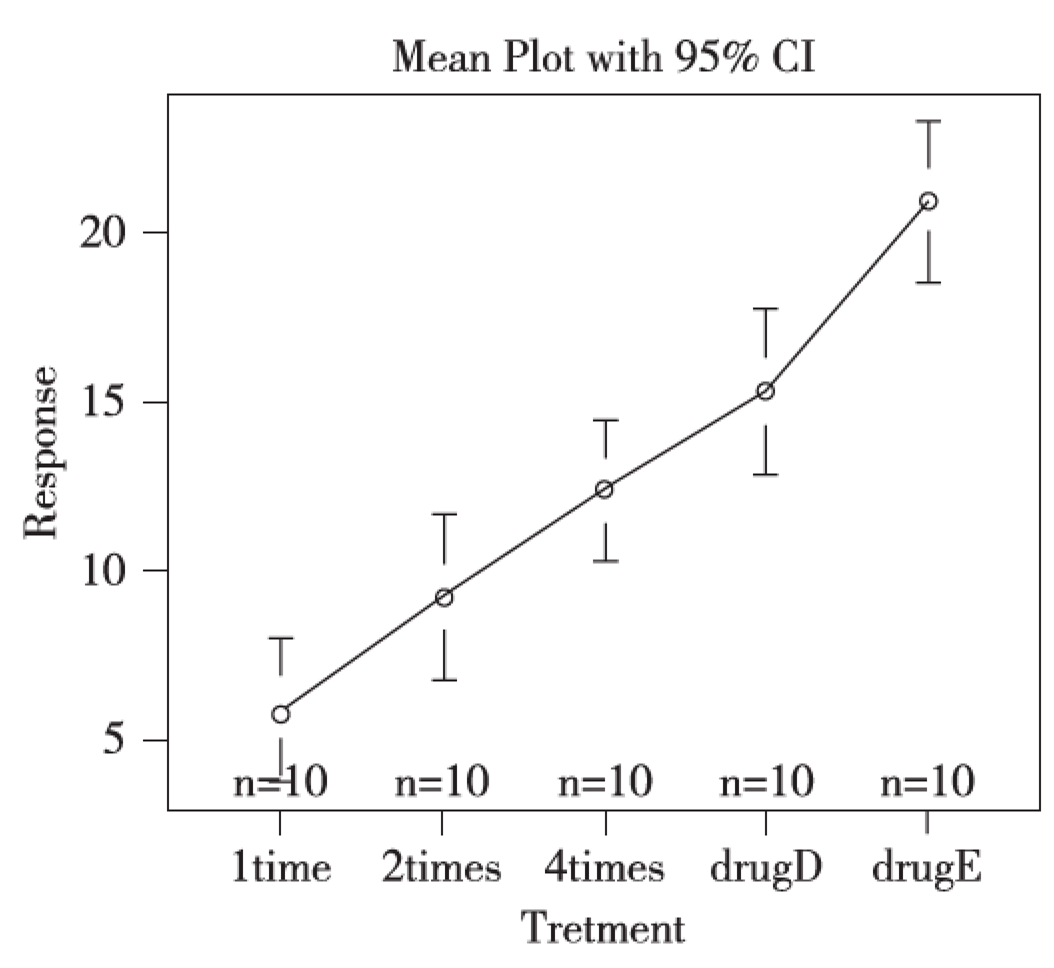


图7.17　五种降低胆固醇药物疗法的均值，95%的置信区间

2. 多重比较

虽然ANOVA对各疗法的F检验表明五种药物疗法效果不同，但是并没有告诉用户哪种疗法与其他疗法不同。多重比较可以解决这个问题。例如，TukeyHSD()函数提供了对各组均值差异的成对检验。

```
> TukeyHSD(fit)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = response ~ trt)

$trt
                  diff        lwr       upr     p adj
2times-1time   3.44300 -0.6582817  7.544282 0.1380949
4times-1time   6.59281  2.4915283 10.694092 0.0003542
drugD-1time    9.57920  5.4779183 13.680482 0.0000003
drugE-1time   15.16555 11.0642683 19.266832 0.0000000
4times-2times  3.14981 -0.9514717  7.251092 0.2050382
drugD-2times   6.13620  2.0349183 10.237482 0.0009611
drugE-2times  11.72255  7.6212683 15.823832 0.0000000
drugD-4times   2.98639 -1.1148917  7.087672 0.2512446
drugE-4times   8.57274  4.4714583 12.674022 0.0000037
drugE-drugD    5.58635  1.4850683  9.687632 0.0030633
```

```
> par(las = 2)
> par(mar = c(5, 8, 4, 2))
> plot(TukeyHSD(fit))
> par(opar)
```

可以看到，1time和2times的均值差异不显著（p=0.138），而1time和4times间的差异非常显著（p<0.001）。成对比较图形如图7.18所示。第一个par语句用来旋转轴标签，第二个用来增大左边界的面积，可使标签摆放更美观。图形中置信区间包含0的疗法说明差异不显著（p>0.05）。

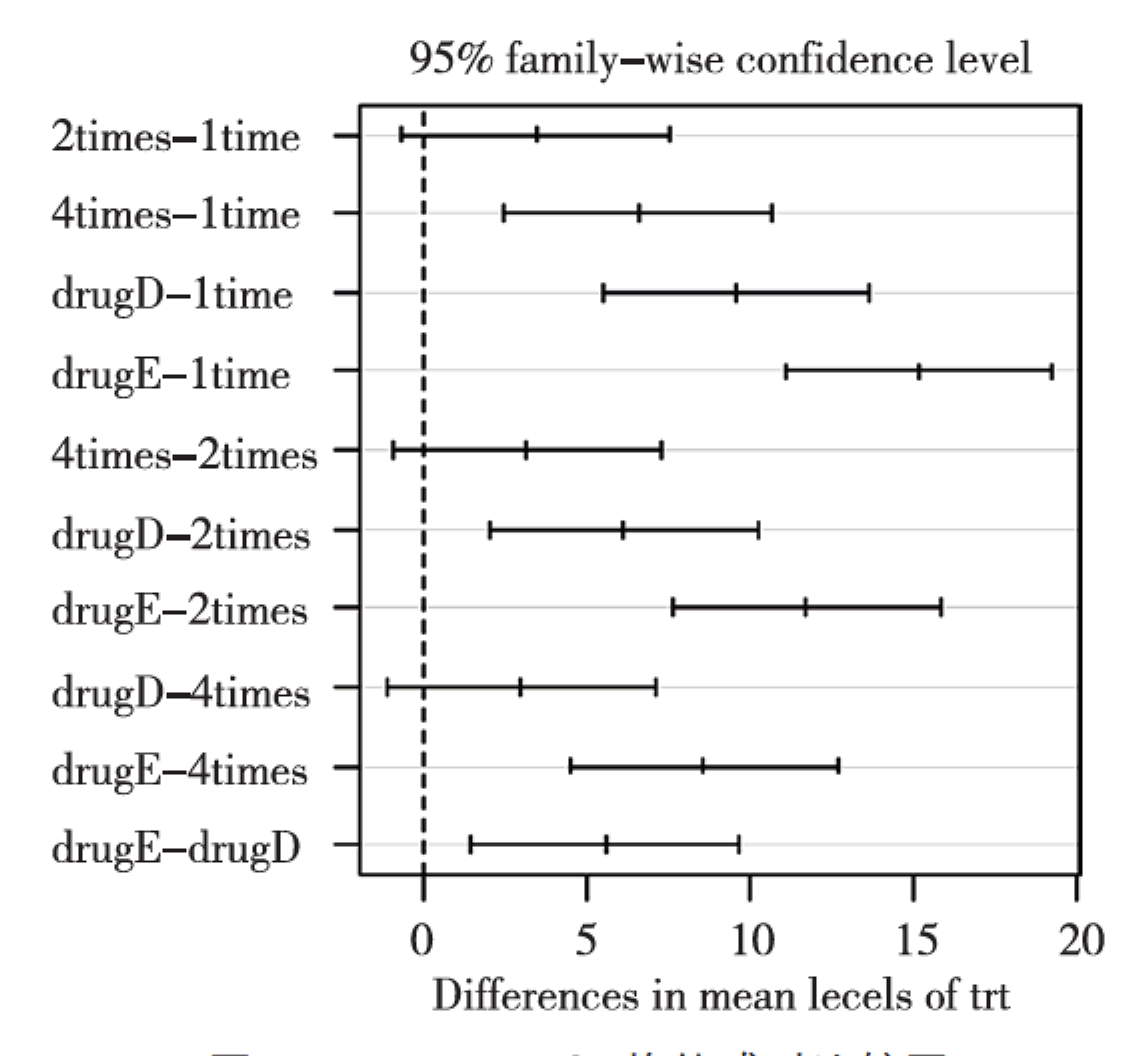


图7.18　Tukey HSD均值成对比较图

3. 双因素方差分析

对于两因素的方差分析，基本思想和方法与单因素的方差分析相似，前提条件仍然是要满足独立、正态、方差齐性。与单因素方差分析所不同的是在双因素方差分析中，有时会出现交互作用，即二因素的不同水平交叉搭配对指标产生影响，这就出现两种情况，有交互作用的双因素分析和无交互作用的双因素分析。在R软件中，方差分析函数aov()既适合于单因素方差分析，也同样适用于双因素方差分析，其中方差模型公式为x ~ A+B，加号表示两个因素具有可加的。在此不做详例说明，可参考专门的R统计书籍。

7.5 R的高级数据分析

R除了上述的初级数据分析方法之外，还有一些较为高级的数据分析方法，例如广义线性模型、聚类分析、判别分析、主成分分析、因子分析等，下面逐一介绍这些高级数据分析方法。

7.5.1 广义线性模型

许多广泛应用的、流行的数据分析方法其实都归属于广义线性模型框架。首先，将简短回顾这些方法背后的理论。现假设想要对响应变量Y和p个预测变量$X_1,\cdots X_p$间的关系进行建模。在标准线性模型中，你可以假设Y呈正态分布，关系的公式为：

$$\mu_Y=\beta_0+\sum_{j=1}^{p}\beta_j X_j \tag{7-2}$$

该等式表明响应变量的条件均值是预测变量的线性组合。参数β_j指一单位X_j的变化造成的Y预期的变化，β_0指当所有预测变量都为0时Y的预期值。对于该等式，你可通俗地理解为：给定一系列X变量的值，赋予X变量合适的权重，然后将它们加起来，便可预测Y观测值分布的均值。

值得注意的是，这里并没有对预测变量X_j做任何分布的假设，与Y不同，它们不需要呈正态分布。实际上，它们常为类别型变量（比如方差分析设计）。另外，对预测变量使用非线性函数也是允许的，比如常会使用预测变量X^2或者$X_1\times X_2$，只要等式的参数（$\beta_0,\beta_1,\ldots,\beta_p$）为线性即可。

广义线性模型拟合的公式为：

$$g(\mu_Y)=\beta_0+\sum_{j=1}^{p}\beta_j X_j \tag{7-3}$$

其中$g(\mu_Y)$是条件均值的函数（称为连接函数）。另外，可放弃Y为正态分布的假设，改为Y服从指数分布族中的一种分布即可。设定好连接函数和概率分布后，便可以通过最大似然估计的多次迭代推导出各参数值。

1. glm()函数

在R中可通过glm函数（还可用其他专门的函数）拟合广义线性模型。它的形式与lm()类

似，只是多了一些参数。函数的基本形式为：

```
glm(formula, family=family(link=function), data= )
```

表7.11列出了概率分布（family）和相应默认的连接函数（function）。

表7.11 glm()的参数

分布族	默认的连接函数
binomial	(link = "logit")
gaussian	(link = "identity")
gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
Quasibinomial	(link = "logit")
quasipoisson	(link = "log")

glm()函数可以拟合许多流行的模型，比如Logistic回归、泊松回归和生存分析（此处不考虑）。下面对Logistic模型进行阐述。假设你有一个响应变量（Y）、三个预测变量（X_1、X_2、X_3）和一个包含数据的数据框（mydata）。

Logistic回归适用于二值响应变量（0,1）。模型假设Y服从二项分布，线性模型的拟合形式为：

$$\ln(\frac{\pi}{1-\pi}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \qquad (7\text{–}4)$$

其中$\pi=\mu_Y$是Y的条件均值（即给定一系列X的值时Y=1的概率）；$\pi/(1-\pi)$为Y=1时的优势比；$\log(\pi/(1-\pi))$为对数优势比，也可省略为logit。本例中$\log(\pi/(1-\pi))$为连接函数，概率分布为二项分布，可用如下代码拟合Logistic回归模型：

```
glm(Y~X1+X2+X3, family=binomial(link="logit"), data=mydata)
```

2. Logistic回归

当通过一系列连续型或类别型预测变量来预测二值型结果变量时，Logistic回归是一个非常有用的工具。以AER包中的数据框Affairs为例，将通过探究婚外情的数据来阐述Logistic回归的过程。首次使用该数据前，请确保已下载和安装此软件包（使用install.packages("AER"))。婚外情数据即著名的“Fair’s Affairs”，取自于1969年《今日心理》（Psychology Today）所做的一个调查，而Greene（2003）和Fair（1978）都对它进行过分析。该数据从601个参与者身上收集了9个变量，包括一年来婚外私通的频率以及参与者性别、年龄、婚龄、是否有小孩、宗教信仰程度（5分制，1分表示反对，5分表示非常信仰）、学历、职业（逆向编号的戈登7种分类），还有对婚姻的自我评分（5分制，1表示非常不幸福，5表示非常幸福）。

先看一些描述性的统计信息：

```
> data(Affairs, package = "AER")
> summary(Affairs)
```

```
   affairs          gender          age          yearsmarried     children
Min.   : 0.000   female:315   Min.   :17.50   Min.   : 0.125   no :171
1st Qu.: 0.000   male  :286   1st Qu.:27.00   1st Qu.: 4.000   yes:430
Median : 0.000                Median :32.00   Median : 7.000
Mean   : 1.456                Mean   :32.49   Mean   : 8.178
3rd Qu.: 0.000                3rd Qu.:37.00   3rd Qu.:15.000
Max.   :12.000                Max.   :57.00   Max.   :15.000
religiousness     education       occupation       rating
Min.   :1.000   Min.   : 9.00   Min.   :1.000   Min.   :1.000
1st Qu.:2.000   1st Qu.:14.00   1st Qu.:3.000   1st Qu.:3.000
Median :3.000   Median :16.00   Median :5.000   Median :4.000
Mean   :3.116   Mean   :16.17   Mean   :4.195   Mean   :3.932
3rd Qu.:4.000   3rd Qu.:18.00   3rd Qu.:6.000   3rd Qu.:5.000
Max.   :5.000   Max.   :20.00   Max.   :7.000   Max.   :5.000
```

```
> table(Affairs$affairs)
```

```
  0   1   2   3   7  12
451  34  17  19  42  38
```

从这些统计信息可以看到，52%的调查对象是女性，72%的人有孩子，样本年龄的中位数为32岁。对于响应变量，72%的调查对象表示过去一年中没有婚外情（451/601），而有婚外情的最多次数为12（占了6%）。

虽然这些婚姻的轻率举动次数被记录下来，但此处我们感兴趣的是二值型结果（有过一次婚外情/没有过婚外情）。按照如下代码，可将affairs转化为二值型因子ynaffair。

```
> Affairs$ynaffair[Affairs$affairs > 0] <- 1
> Affairs$ynaffair[Affairs$affairs == 0] <- 0
> Affairs$ynaffair <- factor(Affairs$ynaffair, levels = c(0, 1), labels = c("No", "Yes"))
> table(Affairs$ynaffair)
```

```
 No Yes
451 150
```

该二值型因子现可作为Logistic回归的结果变量：

```
> fit.full <- glm(ynaffair ~ gender + age + yearsmarried + children + religiousness + education + occupation + rating, data = Affairs, family = binomial())
> summary(fit.full)
```

```
Call:
glm(formula = ynaffair ~ gender + age + yearsmarried + children +
    religiousness + education + occupation + rating, family = binomial(),
    data = Affairs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5713  -0.7499  -0.5690  -0.2539   2.5191

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.37726    0.88776   1.551 0.120807
gendermale     0.28029    0.23909   1.172 0.241083
age           -0.04426    0.01825  -2.425 0.015301 *
yearsmarried   0.09477    0.03221   2.942 0.003262 **
childrenyes    0.39767    0.29151   1.364 0.172508
religiousness -0.32472    0.08975  -3.618 0.000297 ***
education      0.02105    0.05051   0.417 0.676851
occupation     0.03092    0.07178   0.431 0.666630
rating        -0.46845    0.09091  -5.153 2.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 675.38  on 600  degrees of freedom
Residual deviance: 609.51  on 592  degrees of freedom
AIC: 627.51

Number of Fisher Scoring iterations: 4
```

从回归系数的p值（最后一栏）可以看到，性别、是否有孩子、学历和职业对方程的贡献都不显著（你无法拒绝参数为0的假设）。去除这些变量重新拟合模型，检验新模型是否拟合得好：

```
> fit.reduced <- glm(ynaffair ~ age + yearsmarried + religiousness + rating,
data = Affairs, family = binomial())
> summary(fit.reduced)
```

```
Call:
glm(formula = ynaffair ~ age + yearsmarried + religiousness +
    rating, family = binomial(), data = Affairs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6278  -0.7550  -0.5701  -0.2624   2.3998

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.93083    0.61032   3.164 0.001558 **
age           -0.03527    0.01736  -2.032 0.042127 *
yearsmarried   0.10062    0.02921   3.445 0.000571 ***
religiousness -0.32902    0.08945  -3.678 0.000235 ***
rating        -0.46136    0.08884  -5.193 2.06e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 675.38  on 600  degrees of freedom
Residual deviance: 615.36  on 596  degrees of freedom
AIC: 625.36

Number of Fisher Scoring iterations: 4
```

新模型的每个回归系数都非常显著（$p<0.05$）。由于两模型嵌套（fit.reduced是fit.full的一个子集），你可以使用anova()函数对它们进行比较，对于广义线性回归，可用卡方检验。

```
> anova(fit.reduced, fit.full, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: ynaffair ~ age + yearsmarried + religiousness + rating
Model 2: ynaffair ~ gender + age + yearsmarried + children + religiousness +
    education + occupation + rating
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       596     615.36
2       592     609.51  4   5.8474   0.2108
```

结果的卡方值不显著（$p=0.21$），表明四个预测变量的新模型与九个完整预测变量的模型拟合程度一样好。从中可以看出添加性别、孩子、学历和职业变量不会显著提高方程的预测精度，因此可以依据更简单的模型进行解释。

7.5.2 聚类分析

聚类分析是研究“物以类聚”的一种方法，也有人称它为群分析、点群分析、簇群分析等。系统聚类法是将n个样品分成若干类的方法[①]，首先将n个样品各自看成一类，然后计算n类两两间的距离（类之间的距离有多种定义方法），根据结果将距离最近的两类合并成新的一类，接着计算新类与各当前类的距离，再将距离最近的两类合并，这样每次减少一类，直

① http://www.doc88.com/p-865110759511.html

到所有的样品都成为一类[1]。

在R软件中，dist()函数给出了各种距离的计算结果，其调用格式为：

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
```

说明：method表示计算距离的方法，默认值为euclidean（欧氏）距离；diag是逻辑变量，当diag=TRUE时，输出距离矩阵对角线上的距离；upper也是逻辑变量，当upper=TRUE时，输出距离矩阵上三角部分（默认仅输出下三角矩阵）。

定义类与类之间的距离有许多方法，主要有以下七种：

- 类平均法（average Linkage）；
- 重心法（centroid method）；
- 中间距离法（median method）；
- 最长距离法（complete method）；
- 最短距离法（single method）；
- 离差平方和法（ward method）；
- Mcquitty相似法（Mcquitty method）。

各类方法计算方式不同，有学者推荐采用离差平方和法或最短距离法。

利用R语言的hclust()函数就可完成系统聚类分析，其基本调用格式如下：

```
hclust(d, method = "complete", members=NULL)
```

说明：d是由“dist”构成的距离结构，method是系统聚类的方法（默认地是最长距离法）具体说明见R的帮助。

例：设有5个产品，每个产品测得一项质量指标x，其值如下：1，2，4.5，6，8，试用最短距离法、最长距离法、中间距离法、离差平方和法分别对5个产品按质量指标进行分类。

解R程序如下：

```
> x<-c(1, 2, 4.5, 6, 8)
> dim(x)<-c(5, 1)
> d<-dist(x)
> hc1<-hclust(d, "single")
> hc2<-hclust(d, "complete")
> hc3<-hclust(d, "median")
> hc4<-hclust(d, "ward")
> opar<-par(mfrow=c(2, 2))
> plot(hc1, hang=-1);plot(hc2, hang=-1)
> plot(hc3, hang=-1);plot(hc4, hang=-1)
> par(opar)
```

R程序执行的结果见图7.19。可见，四种分类方法结果一致，都将第1，2个分在一类，其余在第二类。

① http://wenku.baidu.com/view/96d4730bba1aa8114431d90c.html

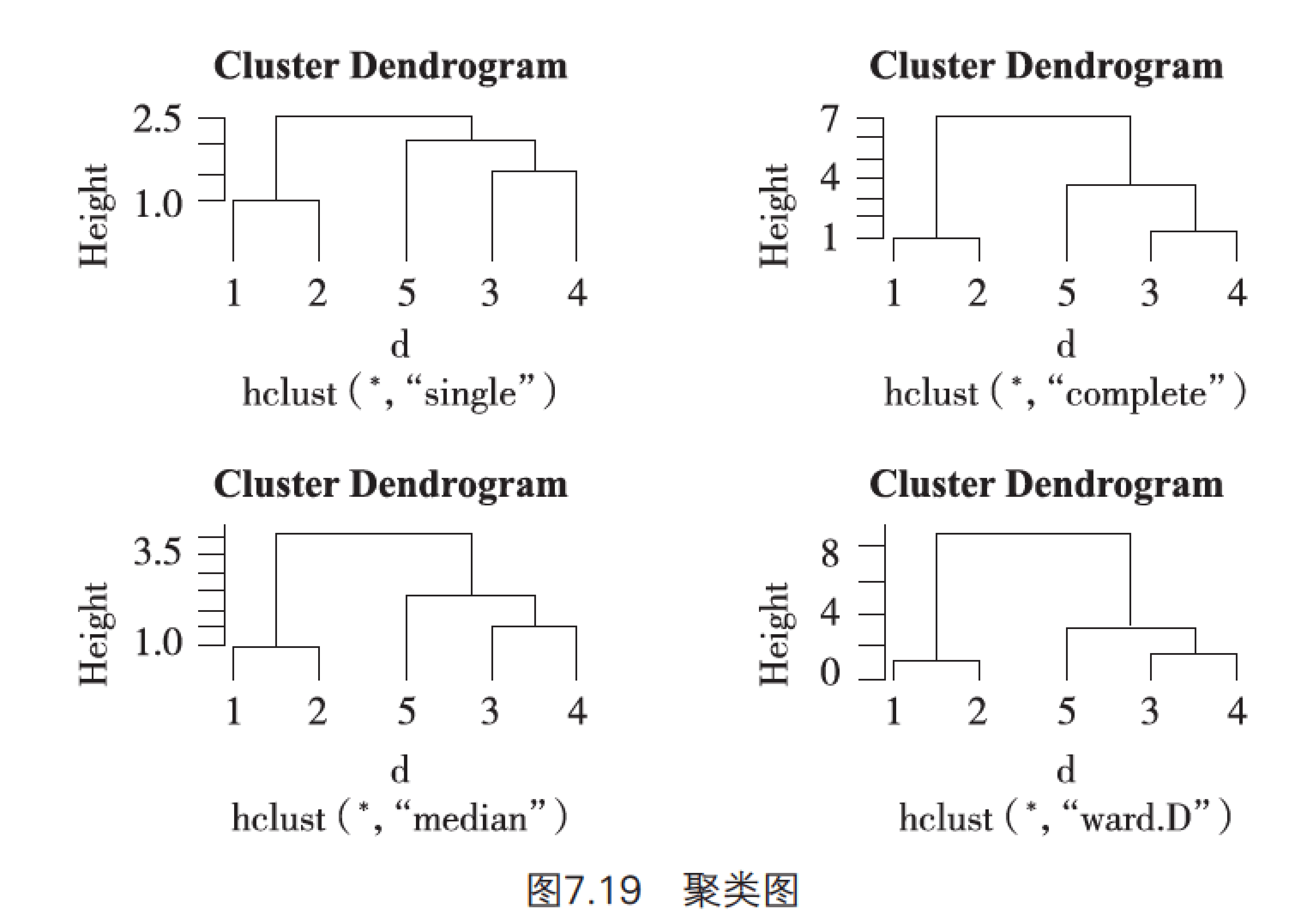


图7.19　聚类图

7.5.3　判别分析

判别分析是一种可用于判断样品所属类型的统计分析方法。判别分析的目的是对已知归类的数据建立由数值指标构成的归类规则，然后把这些规则应用到未知归类的样品中去。常见的判别方法有Fisher判别法和距离判别法等。

Fisher判别法的基本思想[①]是投影，将k组m维数据向某个方向投影，使得投影后组与组之间尽可能地分开（借助于一元方差分析的思想衡量组与组之间是否分开）。距离判别法（或称直观判别法）的基本思想是[②]：根据样品和总体距离的远近，判断属于哪个总体，距离最近的就将它判为属于那个总体。

首先要用命令：

```
>library(MASS)
```

加载MASS宏包，再用函数lda()就可完成Fisher判别分析，其基本调用格式如下：

```
lda(formula, data,…, subset, na.action)
```

说明：formula用法为groups ~ x1+x2+...,group表明总体来源；x1,x2,...表示分类指标；subset指明训练样本。具体说明见R帮助。

Fisher于1936年发表的鸢尾花（Iris）数据被广泛地作为判别分析的例子，数据是对3种（species）鸢尾花：刚毛鸢尾花（setosa）、变色鸢尾花（versicolor）、弗吉尼亚鸢尾花（virginica）各抽取一个容量为50的样本，测量其花萼长（Sepal.Lenth）、花萼宽（Sepal.Width）、花瓣长（Petal.Lenth）、花瓣宽（Petal.Width），单位为mm。试调用R内置文件中的iris数据进行判别分析。

```
> data(iris)
> attach(iris)
> names(iris)
```

① http://wenku.baidu.com/view/652a988071fe910ef12df840.html

② http://wenku.baidu.com/view/ba855cd4360cba1aa811daf0.html

```
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
[5] "Species"
```

```
> iris.lda <- lda(Species ~ Sepal.Length + Sepal.Width+Petal.Length + Petal.Width)
> iris.lda
```

```
Call:
lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

```
> iris.pred=predict(iris.lda) $ class
> table(iris.pred, Species)
```

```
             Species
iris.pred     setosa versicolor virginica
  setosa          50          0         0
  versicolor       0         48         1
  virginica        0          2        49
```

结果说明。

- Group means：包含了每组的平均向量；
- Coefficients of linear discriminants：线性判别系数；
- Proportion of trace：表明了第i判别式对区分各组的贡献大小；
- Species：表明将原始数据代入线性判别函数后的判别结果，setosa组没有错判，versicolor有两个错判，virginica只有一个错判。

7.5.4　主成分分析

PCA的目标是用一组较少的不相关变量代替大量相关变量，同时尽可能保留初始变量的信息，这些推导所得的变量称为主成分，它们是观测变量的线性组合。如第一主成分为：

$$PC_1 = a_1X_1 + a_2X_2 + \ldots + a_kX_k \qquad (7\text{-}5)$$

它是k个观测变量的加权组合，对初始变量集的方差解释性最大。第二主成分也是初始变量的线性组合，对方差的解释性排第二，同时与第一主成分正交（不相关）。后面每一个主成分都最大化它对方差的解释程度，同时与之前所有的主成分都正交。理论上来说，可以选取与变量数相同的主成分，但从实用的角度来看，都希望能用较少的主成分来近似全变量集。下面看一个简单的示例。

数据集USJudgeRatings包含了律师对美国高等法院法官的评分。数据框包含43个观测，12个变量。表7.12列出了所有的变量。

表7.12　USJudgeRatings数据集中的变量

变 量	描 述	变 量	描 述
CONT	律师与法官的接触次数	PREP	审理前的准备工作
INTG	法官正直程度	FAMI	对法律的熟稔程度
DMNR	风度	ORAL	口头裁决的可靠度
DILG	勤勉度	WRIT	书面裁决的可靠度
CFMG	案例流程管理水平	PHYS	体能
DECI	决策效率	RTEN	是否值得保留

（1）判断主成分的个数

以下是一些可用来判断PCA中需要多少个主成分的准则：

- 根据先验经验和理论知识判断主成分数；
- 根据要解释变量方差的积累值的阈值来判断需要的主成分数；
- 通过检查变量间$k \times k$的相关系数矩阵来判断保留的主成分数。

最常见的是基于特征值的方法。每个主成分都与相关系数矩阵的特征值相关联，第一主成分与最大的特征值相关联，第二主成分与第二大的特征值相关联，依此类推。Kaiser–Harris准则建议保留特征值大于1的主成分，特征值小于1的成分所解释的方差比包含在单个变量中的方差更少。Cattell碎石检验则绘制了特征值与主成分数的图形。这类图形可以清晰地展示图形弯曲状况，在图形变化最大处之上的主成分都可保留。最后，你还可以进行模拟，依据与初始矩阵相同大小的随机数据矩阵来判断要提取的特征值。若基于真实数据的某个特征值大于一组随机数据矩阵相应的平均特征值，那么该主成分可以保留。该方法称作平行分析（详见Hayton、Allen和Scarpello的“Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis”，2004）。利用fa.parallel()函数，可以同时对三种特征值判别准则进行评价。对于11种评分（删去了CONT变量），代码如下：

```
>library(psych)
>fa.parallel(USJudgeRatings[, -1], fa = "pc", n.iter=100,show.legend =
FALSE, main = "Scree plot with parallel analysis")
```

代码生成图形见图7.20，展示了基于观测特征值的碎石检验（由线段和x符号组成）、根据100个随机数据矩阵推导出来的特征值均值（虚线），以及大于1的特征值准则（y=1的水平线）。

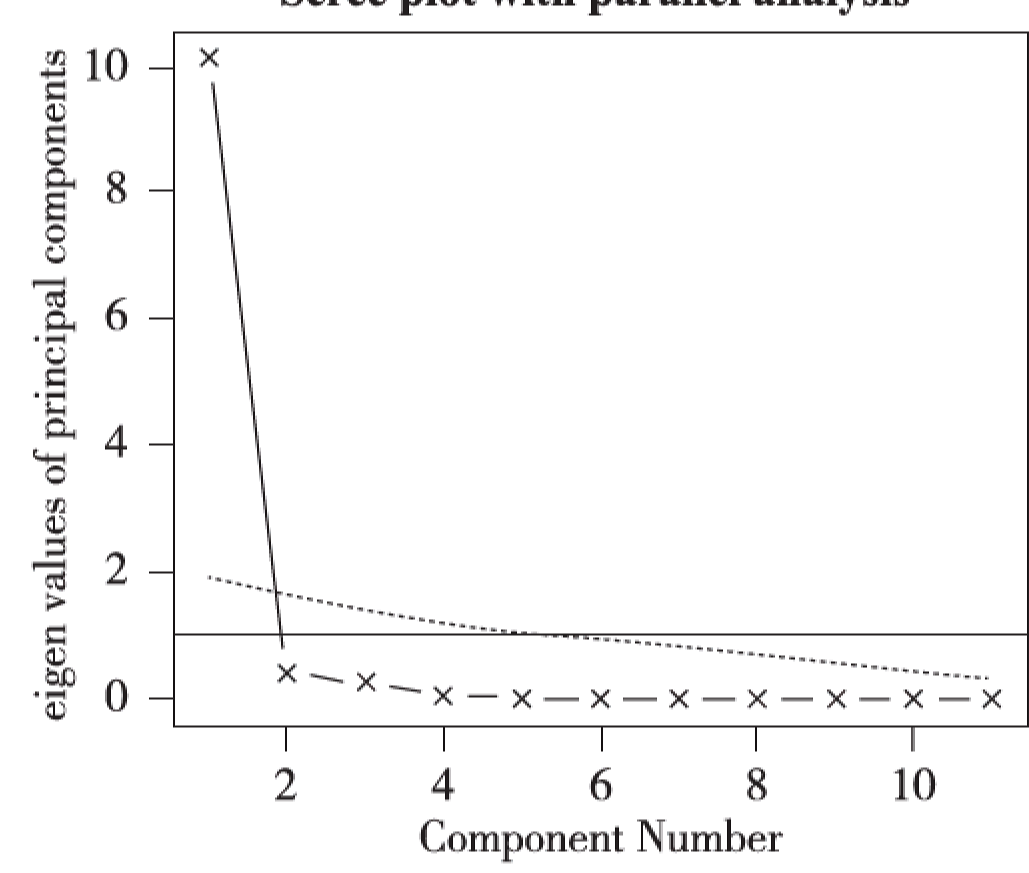


图7.20　评价美国法官评分中要保留的主成分个数

三种准则表明选择一个主成分即可保留数据集的大部分信息。下一步是使用principal()函数挑选出相应的主成分。

（2）提取主成分

之前已经介绍过，principal()函数可以根据原始数据矩阵或者相关系数矩阵做主成分分析。格式为：

```
Principal(r, nfactors=, rotate=, scores=)
```

其中：

- r是相关系数矩阵或原始数据矩阵；
- nfactors设定主成分数（默认为1）；
- rotate指定旋转的方法[默认最大方差旋转（varimax）]；
- scores设定是否需要计算主成分得分（默认不需要）。

使用以下代码可获取第一主成分。

```
> pc <- principal(USJudgeRatings[, -1], nfactors = 1, score = TRUE)
> pc
```

```
Principal Components Analysis
Call: principal(r = USJudgeRatings[, -1], nfactors = 1, scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
      PC1   h2     u2
INTG 0.92 0.84 0.1565
DMNR 0.91 0.83 0.1663
DILG 0.97 0.94 0.0613
CFMG 0.96 0.93 0.0720
DECI 0.96 0.92 0.0763
PREP 0.98 0.97 0.0299
FAMI 0.98 0.95 0.0469
ORAL 1.00 0.99 0.0091
WRIT 0.99 0.98 0.0196
PHYS 0.89 0.80 0.2013
RTEN 0.99 0.97 0.0275

                 PC1
SS loadings    10.13
Proportion Var  0.92
```

此处，输入的是没有CONT变量的原始数据，并指定获取一个未旋转的主成分。由于PCA只对相关系数矩阵进行分析，在获取主成分前，原始数据将会被自动转换为相关系数矩阵。

PC1栏包含了成分载荷，指观测变量与主成分的相关系数。如果提取不止一个主成分，那么还将会有PC2、PC3等栏。成分载荷（component loadings）可用来解释主成分的含义。此处可以看到，第一主成分（PC1）与每个变量都高度相关，也就是说，它是一个可用来进行一般性评价的维度。

h2栏指成分公因子方差——主成分对每个变量的方差解释度。u2栏指成分惟一性——方差无法被主成分解释的比例（1–h2）。例如，体能（PHYS）80%的方差都可用第一主成分来解释，20%不能。相比而言，PHYS是用第一主成分表示性最差的变量。

SS loadings行包含了与主成分相关联的特征值，指的是与特定主成分相关联的标准化后的方差值（本例中，第一主成分的值为10）。最后，Proportion Var行表示的是每个主成分对整个数据集的解释程度。此处可以看到，第一主成分解释了11个变量92%的方差。

7.5.5 因子分析

EFA的目标是通过发掘隐藏在数据下的一组较少的、更为基本的无法观测的变量，来解释一组可观测变量的相关性。这些虚拟的、无法观测的变量称作因子。（每个因子被认为可解释多个观测变量间共有的方差，因此准确来说，它们应该称作公共因子。）

模型的公式为：

$$X_i = a_1F_1 + a_2F_2 + \ldots + a_pF_p + U_i \tag{7-6}$$

其中X_i是第i个可观测变量（i=1,…,k），F_j是公共因子（j=1,…,p），并且$p<k$。U_i是X_i变量独有的部分（无法被公共因子解释）。a_i可认为是每个因子对复合而成的可观测变量的贡献值。

虽然PCA和EFA存在差异，但是它们的许多分析步骤都是相似的。为阐述EFA的分析过

程，我们用它来对六个心理学测验间的相关性进行分析。112个人参与了六个测验，包括非语言的普通智力测验（general）、画图测验（picture）、积木图案测验（blocks）、迷津测验（maze）、阅读测验（reading）和词汇测验（vocab）。我们如何用一组较少的、潜在的心理学因素来解释参与者的测验得分呢?

数据集ability.cov提供了变量的协方差矩阵，可用cov2cor()函数将其转化为相关系数矩阵。数据集没有缺失值。

```
> options(digits = 2)
> covariances <- ability.cov$cov
> correlations <- cov2cor(covariances)
> correlations
```

```
        general picture blocks maze reading vocab
general    1.00    0.47   0.55 0.34    0.58  0.51
picture    0.47    1.00   0.57 0.19    0.26  0.24
blocks     0.55    0.57   1.00 0.45    0.35  0.36
maze       0.34    0.19   0.45 1.00    0.18  0.22
reading    0.58    0.26   0.35 0.18    1.00  0.79
vocab      0.51    0.24   0.36 0.22    0.79  1.00
```

因为要寻求用来解释数据的潜在结构，可使用EFA方法。与使用PCA相同，下一步工作为判断需要提取几个因子。

（1）判断需提取的公共因子数

用fa.parallel()函数可判断需提取的因子数：

```
> library(psych)
> covariances <- ability.cov$cov
> correlations <- cov2cor(covariances)
> fa.parallel(correlations, n.obs = 112, fa = "both", n.iter = 100, main =
"Scree plots with parallel analysis")
```

结果见图7.21。注意，代码中使用了fa="both"，因子图形将会同时展示主成分和公共因子分析的结果。

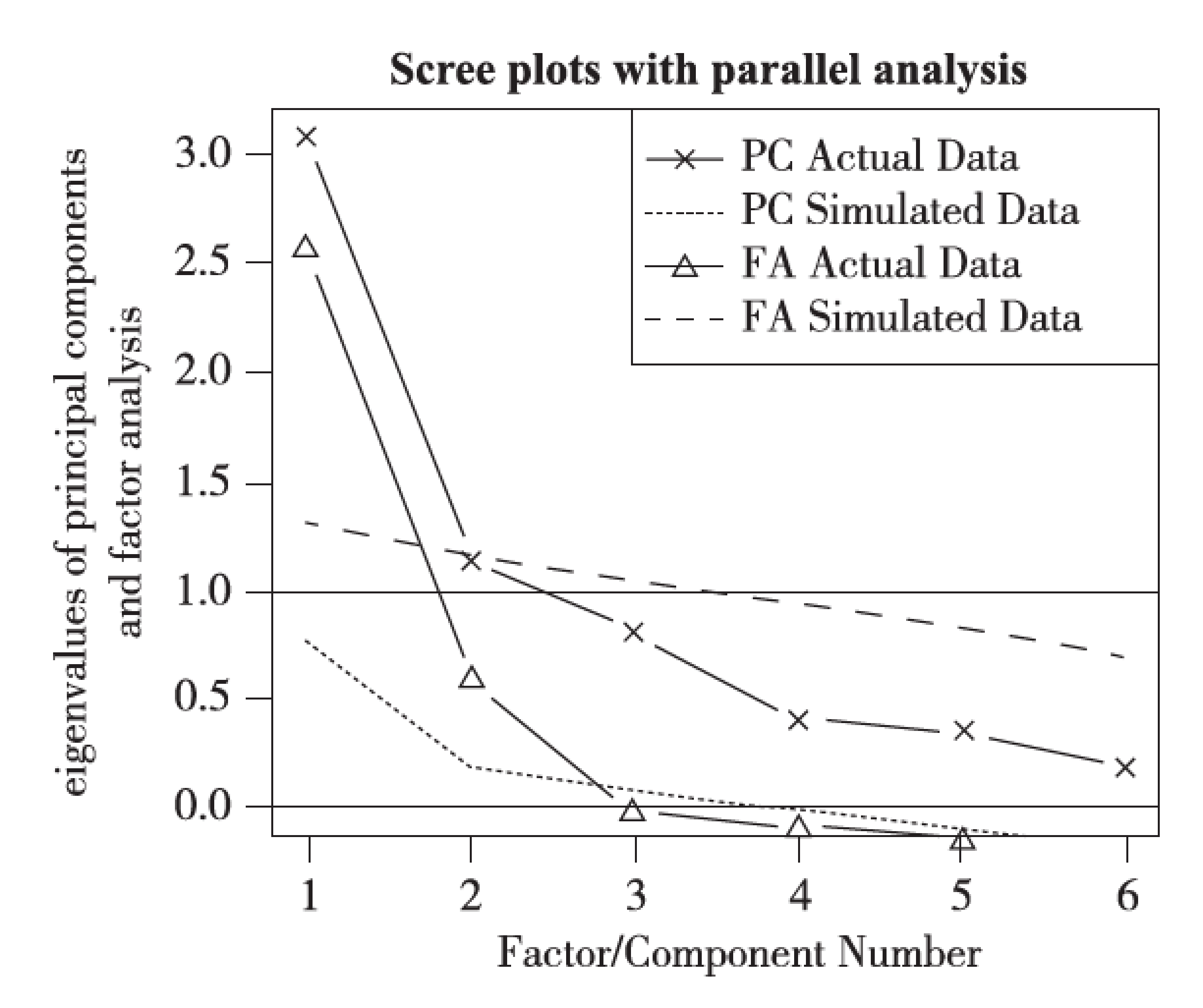


图7.21　判断心理学测验需要保留的因子数

在图7.21中有几个值得注意的地方。如果使用PCA方法，我们可能会选择一个成分（碎石检验和平行分析）或者两个成分（特征值大于1）。当摇摆不定时，高估因子数通常比低估因子数的结果好，因为高估因子数一般较少曲解“真实”情况。

观察EFA的结果，显然需提取两个因子。碎石检验的前两个特征值（三角形）都在拐角处之上，并且大于基于100次模拟数据矩阵的特征值均值。对于EFA，Kaiser–Harris准则的特征值数大于0，而不是1。（大部分人都没有意识到这一点。）在图7.21中该准则也建议选择两个因子。

（2）提取公共因子

现在决定提取两个因子，可以使用fa()函数获得相应的结果。fa()函数的格式如下：

```
fa(r, nfactors=, n.obs=, rotate=, scores=, fm=)
```

其中：

- r是相关系数矩阵或者原始数据矩阵；
- nfactors设定提取的因子数（默认为1）；
- n.obs是观测数（输入相关系数矩阵时需要填写）；
- rotate设定旋转的方法（默认互变异数最小法）；
- scores设定是否计算因子得分（默认不计算）；
- fm设定因子化方法（默认极小残差法）。

与PCA不同，提取公共因子的方法很多：最大似然法（ml）、主轴迭代法（pa）、加权最小二乘法（wls）、广义加权最小二乘法（gls）和最小残差法（minres）。统计学家青睐使用最大似然法，因为它有良好的统计性质。不过有时候最大似然法不会收敛，此时使用主轴迭代法效果会很好。

本例使用主轴迭代法（fm = "pa"）提取未旋转的因子。执行下述代码后，结果为：

```
> fa <- fa(correlations, nfactors = 2, rotate = "none", fm = "pa")
> fa
```

```
Factor Analysis using method =  pa
Call: fa(r = correlations, nfactors = 2, rotate = "none", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
         PA1   PA2   h2    u2 com
general 0.75  0.07 0.57 0.432 1.0
picture 0.52  0.32 0.38 0.623 1.7
blocks  0.75  0.52 0.83 0.166 1.8
maze    0.39  0.22 0.20 0.798 1.6
reading 0.81 -0.51 0.91 0.089 1.7
vocab   0.73 -0.39 0.69 0.313 1.5

                       PA1  PA2
SS loadings           2.75 0.83
Proportion Var        0.46 0.14
Cumulative Var        0.46 0.60
```

可以看到，两个因子解释了六个心理学测验60%的方差。不过因子载荷阵的意义并不太好解释，此时使用因子旋转将有助于因子的解释。

（3）因子旋转

可以使用正交旋转或者斜交旋转来旋转上节中两个因子的结果。现在同时尝试下两种方法，看看它们的异同。首先使用正交旋转：

```
> fa.varimax <- fa(correlations, nfactors = 2, rotate = "varimax",  fm =
"pa")
```

```
> fa.varimax
```

```
Factor Analysis using method =  pa
Call: fa(r = correlations, nfactors = 2, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
         PA1  PA2   h2    u2 com
general 0.49 0.57 0.57 0.432 2.0
picture 0.16 0.59 0.38 0.623 1.1
blocks  0.18 0.89 0.83 0.166 1.1
maze    0.13 0.43 0.20 0.798 1.2
reading 0.93 0.20 0.91 0.089 1.1
vocab   0.80 0.23 0.69 0.313 1.2

                       PA1  PA2
SS loadings           1.83 1.75
Proportion Var        0.30 0.29
Cumulative Var        0.30 0.60
```

结果显示因子变得更好解释了。词汇和阅读这两个在第一因子上载荷比较大，积木图案、画图和迷宫这三个在第二因子上载荷较大，非语言的普通智力测量在两个因子上载荷较为平均，这表明存在一个语言智力因子和一个非语言智力因子。使用正交旋转将人为地强制两个因子不相关。如果想允许两个因子相关该怎么办呢？此时可以使用斜交转轴法，比如promax用斜交旋转提取因子：

```
> fa.promax <- fa(correlations, nfactors = 2, rotate = "promax",  fm = "pa")
> fa.promax
```

```
Factor Analysis using method =  pa
Call: fa(r = correlations, nfactors = 2, rotate = "promax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
          PA1   PA2   h2    u2 com
general  0.36  0.49 0.57 0.432 1.8
picture -0.04  0.64 0.38 0.623 1.0
blocks  -0.12  0.98 0.83 0.166 1.0
maze    -0.01  0.45 0.20 0.798 1.0
reading  1.01 -0.11 0.91 0.089 1.0
vocab    0.84 -0.02 0.69 0.313 1.0

                       PA1  PA2
SS loadings           1.82 1.76
Proportion Var        0.30 0.29
Cumulative Var        0.30 0.60
Proportion Explained  0.51 0.49
Cumulative Proportion 0.51 1.00

 With factor correlations of
     PA1  PA2
PA1 1.00 0.57
PA2 0.57 1.00
```

根据以上结果，可以看出正交旋转和斜交旋转的不同之处。对于正交旋转，因子分析的重点在于因子结构矩阵（变量与因子的相关系数）。而对于斜交旋转，因子分析会考虑三个矩阵：因子结构矩阵、因子模式矩阵和因子关联矩阵。

因子模式矩阵即标准化的回归系数矩阵。它列出了因子预测变量的权重。因子关联矩阵即因子相关系数矩阵。

在用promax提取因子时，PA1和PA2栏中的值组成了因子模式矩阵。它们是标准化的回归系数，而不是相关系数。注意，矩阵的列仍用来对因子进行命名（虽然此处存在一些争论）。同样可以得到一个语言因子和一个非语言因子。

因子关联矩阵显示两个因子的相关系数为0.57，相关性很大。如果因子间的关联性很低，可能需要重新使用正交旋转来简化问题。

因子结构矩阵（或称因子载荷阵）没有被列出来，但可以使用公式F=P*Phi很轻松地得到它，其中F是因子载荷阵，P为因子模式矩阵，Phi为因子关联矩阵。下面的函数即可进行该

乘法运算：

```
> fsm <- function(oblique) {
+     if (class(oblique)[2]=="fa" & is.null(oblique$Phi)) {
+         warning("Object doesn't look like oblique EFA")
+     } else {
+         P <- unclass(oblique$loading)
+         F <- P %*% oblique$Phi
+         colnames(F) <- c("PA1", "PA2")
+         return(F)
+     }
+ }
```

对上面的例子使用fsm函数，可以得到：

```
> fsm(fa.promax)
```

```
         PA1  PA2
general 0.64 0.69
picture 0.33 0.61
blocks  0.44 0.91
maze    0.25 0.45
reading 0.95 0.47
vocab   0.83 0.46
```

现在可以看到变量与因子间的相关系数。将它们与正交旋转所得因子载荷阵相比，会发现该载荷阵列的噪音比较大，这是因为之前允许潜在因子相关。虽然斜交方法更为复杂，但模型将更符合真实数据。使用factor.plot()或fa.diagram()函数，可以绘制正交或者斜交的结果，如图7.22。来看以下代码：

```
> factor.plot(fa.promax, labels = rownames(fa.promax$loadings))
```

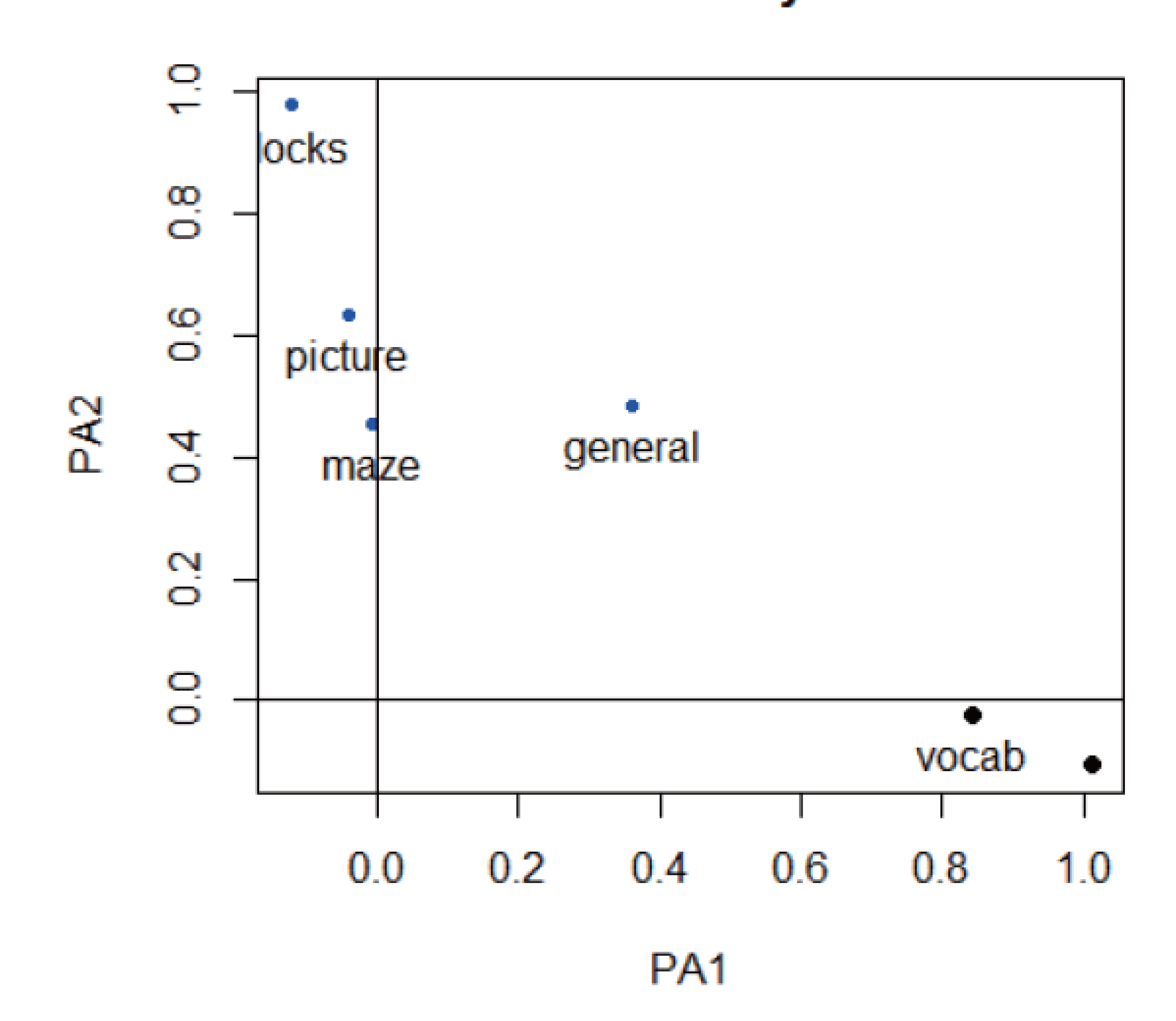


图7.22　数据集ability.cov中心理学测验的两因子图形

使用以下代码：

```
fa.diagram(fa.promax, simple=FALSE)
```

生成的图形如图7.23所示。若使simple=TRUE，那么将仅显示每个因子下最大的载荷，以及因子间的相关系数。这类图形在有多个因子时十分实用。

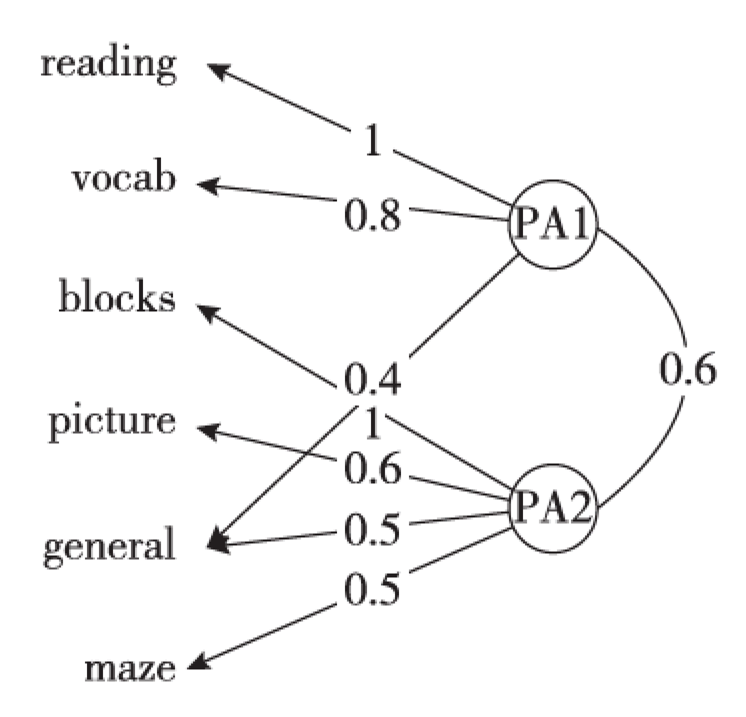


图7.23 数据集ability.cov中心理学测验的两因子斜交旋转结果图

（4）因子得分

相比PCA，EFA并不那么关注计算因子得分。在fa()函数中添加score=TRUE选项（原始数据可得时）便可很轻松地获得因子得分。另外还可以得到得分系数（标准化的回归权重），它在返回对象的weights元素中。

对于ability.cov数据集，通过二因子斜交旋转法便可获得用来计算因子得分的权重：

```
>fa.promax$weights
```

```
         [,1]  [,2]
general 0.080 0.210
picture 0.021 0.090
blocks  0.044 0.695
maze    0.027 0.035
reading 0.739 0.044
vocab   0.176 0.039
```

与可精确计算的主成分得分不同，因子得分只是估计得到的。它的估计方法有多种，fa()函数使用的是回归方法。

7.6 R在大数据处理中的应用

R是一个用于统计计算和统计制图的优秀工具，具有Unix、Linux、MacOS和Windows版本，均可以免费下载以及使用。R强大的功能以及开放性、灵活性使得其在大数据处理中的应用受到广大数据处理、分析人员的喜爱，R已经成为大数据处理应用中一款举足轻重的软件。下面从R处理大数据以及R与Hadoop交互两方面对R在大数据处理中的应用加以介绍。

7.6.1 R处理大数据

R将所有的对象都存储在虚拟内存中。对于大部分人而言，这种设计可以带来很好的交互体验，但如果要处理大量的数据，这就会影响程序的运行速度，带来和内存相关的错误。

具体的内存限制取决于R的版本（32位或64位）和所使用的操作系统。比如，以cannot allocate vector of size开头的错误信息通常都是因为无法获得足够的连续内存空间，以cannot allocate vector of length开头的错误信息表示超过了内存地址的限制。在处理大量的数据时，应该尽可能地用64位版。无论是什么版本，单个向量中元素数量的上限都是2 147 483 647（详见?Memory命令的帮助）。

在处理大数据时，要考虑三个问题：高效执行的程序；将数据保存到外部避免内存问题；用有针对性的统计方法高效地分析海量数据。

1. 高效的程序设计

下面是在处理大型数据时有助于提升性能的程序设计建议。

- 尽可能地做向量化计算。用R内建的函数来处理向量、矩阵和列表，而且要尽量避免使用循环。
- 用矩阵，而不是数据框。
- 在使用read.table()系列函数将外部数据读取到数据框中时，明确地指定colClasses和nrows，设置comment.char = ""，并且用"NULL"标明不需要的列。这样可降低内存的使用量，显著地提高处理速度。在将外部数据读入矩阵时，可以用scan()函数。
- 在完整的数据集上运行程序之前，请先用数据的子集测试程序，以便优化代码并消除bug。
- 删除临时对象和不再需要的对象。调用rm(list=ls())会从内存中删除所有的对象，得到一个干净的环境。要删除特定的对象，可以用rm(object)。

在处理大数据时，提高代码性能也就只能这样。在遇到内存限制时，还可以将数据保存到外部存储器中，并使用特殊的分析方法。

2. 在内存、外存中存储数据

有好几个包可以将数据保存到R的主内存之外。主要的方法是将数据保存到外部数据库中，或是硬盘上的二进制文件中，然后再按需要访问其中的某个部分。表7.13中列出了一些有用的包。

表7.13　用于访问大型数据集的R包

包	描　述
ff	提供了一种数据结构，可以将数据保存到硬盘上，但用起来却像是在内存中
bigmemory	支持大型矩阵的创建、存储、访问和操作。矩阵可以分配在共享内存和内存映射文件中
filehash	实现了一个简单的key-value数据库，用字符串的键值关联到硬盘上存储的数据值
ncdf、ncdf4	提供了Unidate netCDF数据文件的接口
RODBC、RMySQL、ROracle、RPostgreSQL、RSQLite	这些包每一个都可用于访问相应的外部关系型数据库管理系统

上面介绍的这些包都可用于解决R在保存数据时的内存限制问题。不过，在分析大数据时，还需要专门的方法以在可接受的时间内完成分析。

3. 用于大数据的分析包

R有如下几个用于分析大型数据的包。

- biglm和speedglm包能以内存高效的方式实现大型数据的线性模型拟合和广义线性模型拟合。
- 有好几个包是用来分析bigmemory包生成的大型矩阵的。biganalytics包提供了k均值聚类、列统计和一个biglm的封装；bigtabulate包提供了table()、split()和tapply()功能；bigalgebra包提供了高级的线性代数函数。
- biglars包跟ff配合使用，为在内存中无法放置的大数据提供了最小角回归（least-angleregression）、lasso和逐步回归分析。Borbdingnag包可以处理大数字（大于2的1024次方的数）。

在任何编程语言中，处理GB级和TB级的数据都是挑战。关于R中这方面方法的更多信息，可以查看CRAN上的这个Task View：High-Performance and Parallel Computing with R（cran.r-projectorg/web/views/）。

7.6.2 R与Hadoop交互

R会把所有的对象读存入虚拟内存中，在没有遇到大数据集的情况下，这里的大数据集中大数据的数据容量通常是10～100GB或更多，且数据种类包括结构化数据、非结构化数据等多种类型，此时这种设计可以提高与R交互的速度。但在遇到大数据集的情况下，这种设计会使程序运行速度降低，甚至还会产生与内存相关的错误。即便使用64位版的R，向量中的元素个数最大也只能是2 147 483 647。使用R语言处理大数据时，需要采用特殊的方式。

目前，通常有两种方法将R语言与大数据处理平台相结合使用。第一种方法是，在Hadoop上用MapReduce处理PB、TB量级的数据，缩小数据容量到GB量级，然后将其加载到R中进行处理。在R中，GB级别的数据可以利用MPI并行处理框架构建的集群计算。在Rmpi包的基础上可以实现各种MPI支持的并行编程模式。而简单易用snow包及其简化包装版snowfall包则在支持协议的多样性上更好。其中，snow支持SOCKET、MPI、PVM、NWS四种线程通信协议，如果对MPI了解程度不高或者没有安装，也可以直接使用SOCKET方式快速上手。而简化包装版的snowfall包则使得并行化的计算如同普通编程一样简单。由于这些包是为R而扩展出来的，所以跟R的矢量式编程思想能无缝结合，用户只需要将程序用矢量化语言描述出来（比如R的apply系列函数或简单矩阵运算），再移植到snowfall并行计算平台上几乎就是零成本。第二种方法是，直接使用支持Hadoop的R包，在R中操作存放在HDFS中的数据，并利用R语言完成MapReduce算法，用来替代Java的MapReduce实现。RHadoop包使得R语言具有处理高达TB甚至PB级的大数据的能力。在GitHub社区可以找到该项目与开源实现代码。RHadoop包含三个R包，分别是rhdfs，rmr以及rHBase，分别对应Hadoop系统架构中的HDFS，MapReduce和HBase三个部分。除RHadoop包之外，还有从R中进行Hive查询的RHive包，能够直接从Hive中进行查询。

将R和Hadoop结合起来，其既能够利用Hadoop分布式MapReduce计算打破数据量的限制，又能够利用R中的众多优秀的扩展包，快速实现所需的数据处理和分析，是一个值得推荐的大数据处理的最佳实践。

7.7　练习

1. 根据表7.14提供的经济数据，

（1）试画出散点图，判断国民收入（Y）与消费量（X）是否有线性关系；

（2）求出Y关于X的一元线性回归方程；

（3）对方程作显著性检验；

（4）现测得1981年消费量X=3441，试给出1981年国民收入的预测值及相应的区间估计（α =0.05）。

表7.14　我国钢材消费量及国民收入

年份	钢材消费量（万吨）	国民收入（亿元）	年份	钢材消费量（万吨）	国民收入（亿元）
1964	698	1097	1973	1765	2286
1965	872	1284	1974	1762	2311
1966	988	1502	1975	1960	2003
1967	807	1394	1976	1902	2435
1968	738	1303	1977	2013	2625
1969	1025	1555	1978	2446	2948
1970	1316	1917	1979	2736	3155
1971	1539	2051	1980	2825	3372
1972	1561	2111			

2.胃癌的鉴别：表7.15是从病例中随机抽取的部分资料。这里有3个类别（group）：胃癌（ca）、萎缩性胃炎（ga）和非胃炎患者（non）。从每个总体抽5个病人，每人化验4项生化指标：血清铜蛋白（X_1）、蓝色反应（X_2）、尿乙酸（X_3）和中性硫化物（X_4）。试对胃癌检验的生化指标值用Fisher 判别的方法进行判别归类。

表7.15　胃癌检验的生化指标值

类别	序号	血清铜蛋白X_1	蓝色反应X_2	尿乙酸X_3	中性硫化物X_4
胃癌患者	1	228	134	20	11
	2	245	134	10	40
	3	200	167	12	27
	4	170	150	7	8
	5	100	167	20	14
萎缩性胃炎患者	6	225	125	7	14
	7	130	100	6	12
	8	150	117	7	6
	9	120	133	10	26
	10	160	100	5	10

（续表）

类别	序号	血清铜蛋白X_1	蓝色反应X_2	尿乙酸X_3	中性硫化物X_4
非胃炎患者	11	185	115	5	19
	12	170	125	6	4
	13	165	142	5	3
	14	135	108	2	12
	15	100	117	7	2

参考文献

[1] Robert I Kabacoff. R语言实战[M]. 高涛，肖楠，陈钢译. 北京：人民邮电出版社，2013.

[2] 汤银才. R语言与统计分析[M]. 北京：高等教育出版社，2008.

[3] 杨霞，吴东伟. R语言在大数据处理中的应用[J]. 科技咨询. 2013.

[4] 吴喜之. 统计学：从数据到结论（第三版）[M]. 北京：中国统计出版社，2009.

[5] 赵小永，赵政文. 相关性在情感分析上的应用[J]. 开发应用，2011，27（12）：39.

[6] 唐国华. 企业文化建设对企业发展影响的实证检验—以电器行业上市公司为例[J]. 生产研究. 2009，17：168.

[7] 王献勇，刘树惠，赵攀. 判别分析在矿体含矿性判别过程中的应用[J]. 工程地质计算机应用. 2007，4：21.

第8章

大数据用于预测和决策

21世纪是数据化的时代，所有企业都面临着共同的机遇与挑战：如何才能将海量的数据转化为财富和客户价值？另一个重要问题是，企业如何利用这些数据激发下一波的业务创新？而数据的真正价值来自于有效使用数据而作出的决策。企业的经营决策多达数百万个，这些决策对企业与客户的关系以及企业的发展前景会产生影响。高端先进的大数据技术与手段使得人们能够进行准确地分析、预测，确保公司的决策能应对日益增加的复杂性，跟上日益加快的步伐。

8.1 利用分析技术作决策的发展历史和展望

8.1.1 利用分析技术作决策的发展历程

决策是人们为了实现特定的目标，根据客观的可能性，在占有一定信息和经验的基础上，借助一定的工具、技巧和方法，对影响目标实现的诸因素进行分析、计算、判断和选优后，对未来的行动作出决定。决策是人们在政治、经济、技术和日常生活中普遍存在的一种行为，也是管理中经常发生的一种活动。

在决策过程中，决策者面临的主要问题是：对决策问题中的风险进行科学分析并采取有效的方法来降低或消除这些风险；对冲突的多种目标进行科学全面的权衡，从可行的方案中选出满意的决定。信息是人们用来克服风险的一种资源，夺取机遇是决策者的目的，相关资源是实现决策的物质保证。因此，信息、机遇与资源是决策过程的三大要素。决策支持系统的中心任务就是协助决策者统筹与协调好这三要素间的关系。故而，决策分析的科学性与决策模式的正确性是决策成功的关键。

早在20世纪50、60年代，计算机就开始被一些佼佼者用来改善决策了。其中杰出的计算机学家杰伊·弗莱斯特在剑桥提倡用计算机引导商业管理系统，他开发了一种基于计算机模拟的分析和解决问题的方法，这个名为系统动力学的方法能够帮助管理者理解业务流程与决策关系。1961年弗莱斯特针对系统动力学出版了一本《工业动力学》，该书对系统动力学的发展具有巨大影响。20世纪70年代，弗莱斯特在麻省理工学院的学生德内拉·梅多斯应用系统动力学理论设计了一个全球模型，并以此为基础编写了《增长的极限》一书来预测人口增长、经济学和地球环境的发展趋势。

小托马斯·沃森（Thomas Watson Jr.，1914—1993），IBM（国际商用机器公司）的开拓者，有史以来最伟大的资本家，于1952年聘请工程师研制出IBM的第一台可存储程序的计算机IBM 701。这个将IBM的商业机器远景定位为计算机的决策启动了商业信息技术革命，同时开启了大型企业决策管理的先河。

同一时期，在加利福尼亚州，工程师威廉·费尔和数学家厄尔·艾萨克这两位优秀的流程管理学家，在运筹学领域开始了自己的生涯。1956年，两人创立了费埃哲公司，他们认为一个志在卓越的组织，它的经营管理决策应该是有条理的和以数据驱动的，而不是仅仅遵循于直觉和共识。他们的理想是建立供企业使用的、基于计算机的数学工具，加强经营决策，使流程管理成为实现更好的经营绩效的基础。在费埃哲成立3年后，他们开始采用消费信贷来验证他们的想法。随着社会上经营活动越来越计算机化，信用卡的使用日益广泛，这也使得企业开始搜集到有关客户行为的数据。费埃哲推出了第一个用于信用评分的模型，这个模型根据历史数据来分析客户以往的行为，从而预测他们的信用，并依据个人银行存款余额及付款记录等相关变量的分析，得出信用评分来预测人们的还贷能力，这种方法得出的结论远远胜于银行家作出的决策。

20世纪80年代初，美国旧金山大学的管理学教授韦里克提出了SWOT模型，即态势分析法。该方法经常被用于企业战略制定、竞争对手分析等场合[①]。来自于麦肯锡咨询公司的SWOT分析，包括分析企业的优势（Strength）、机会（Opportunity）、劣势（Weakness）与威胁（Threats）。所以，实际上SWOT分析是综合和概括企业内、外部条件各方面的内容，进而对组织面临的机会、优劣势和威胁进行分析的一种方法。SWOT分析能够帮助企业把行动和资源聚集在自己的强项以及有最多机会的地方。除了SWOT模型之外，可同时用于战略分析与战略决策的还有其他模型，例如GE矩阵、波士顿矩阵等。企业经营成败的关键是战略决策，它关系到企业的生存与发展。另外，企业中管理决策、业务决策也同样是企业不可或缺的部分。事实证明，正确的决策能够使企业往正确的方向前进，提高竞争力和适应环境的能力，取得良好的经济效益。反之，失误的决策则会给企业带来巨大的损失，严重者甚至会导致企业破产。

时至21世纪，云计算、社会化媒体以及信息爆炸时代产生的海量数据开始为企业战略决策提供了科学依据，体现在精英智慧、依靠经验判断、自上而下的传统战略论开始走向大数据决策。依据大数据进行决策，从数据中获取价值，让数据主导决策，是一种前所未有的决策方式，并正在推动着人类信息管理准则的重新定位。随着大数据分析和预测性分析对管理决策影响力的逐渐加大，依靠直觉做出决定的状况将会被彻底地改变。

澳大利亚电力集团Energex能够预测各地电力需求，从而确定应该在何处建设电网；联合爱迪生电力公司则能预测在用电高峰时可能出现的系统故障。大数据预测在教育领域也在发挥作用，如美国公立大学系统可预测学生的辍学率，并根据其预测结果来积极管理学生，以降低辍学率。事实上，亚利桑那州立大学、亚拉巴马大学、爱因霍芬科技大学、艾奥瓦州立大学等都在用计算机预测学生的辍学率。目前一些企业已经成功地商业化运作教育中的大数据，在高等教育领域建立起最大的跨校学习数据库。通过这些海量数据，能够看到学生的分

① http://baike.baidu.com/view/147311.htm?func=retitle.

数、出勤率、辍学率和保留率的主要趋势。通过使用100多万名学生的相关记录和700万个课程记录，这些公司的软件能够让用户探测性地知道导致辍学和学习成绩表现不良的警告性信号。此外，该软件还允许用户发现那些导致无谓消耗的特定课程，并且看出哪些资源和干预是最成功的。

当然大数据的作用远远不止这一点，在医疗、交通、能源、材料、商业和服务等行业领域，甚至在新闻传媒领域，也都在以大数据为发展契机，大数据对于管理者的决策有着重大的参考价值。企业是经济系统的发动机。积极地拥抱大数据技术变革，改善企业决策水平，是时代发展的要求，是社会进步的要求。根据戴尔公司（Dell）和微软公司（Microsoft）的调查，美国有超过60%的企业认为需要使用大数据技术，而接近一半的企业打算在未来增加大数据方面的预算。大数据决策越来越受到各行各业的关注。第25届全国医药经济信息发布会以“洞见·预测——信息让决策更优”为主题拉开了序幕。论坛上，标点信息公司与嘉宾们分享了25年来对医药市场研究的专业积累、精选经典案例，共同挖掘信息研究对企业新药研发立项和营销业绩提升的正能量。会上嘉宾认为，通过对产品信息、销售数据等方面的信息进行量化分析，能帮助企业看得更细、更远。中国电子信息产业发展研究院云计算产业研究中心总经理吴李知在《哈佛商业评论》上发表了文章，介绍了企业决策者如何收集数据和利用大数据作决策的方法。在互联网行业和金融行业可以运用大数据对客户的信用风险进行鉴别。大数据还能够预测民航领域诸如机票打折、班机延误等信息；帮助纽约市政府找出发生火灾和井盖爆炸概率较高的地点；帮助快递企业确定合适的行驶路线，从而减少等候的时间；Zynga利用数据分析帮助其修改游戏产品；商场采用大数据分析产品之间存在的关联性……这一切都对我们的决策、行为、习惯产生着影响。

毫无疑问，大数据决策正在以不可阻挡之势扑面而来。

8.1.2　大数据决策的展望

近年来，云计算技术迅速发展，移动互联网、物联网应用大规模爆发，那些由社交媒体、视频、音频、邮件、文档信息和网页所产生的海量数据以惊人的速度在增长。据相关权威机构预测，过去两年生成的数据占整个人类历史数据总量的90%，而全球数据总量每过两年就会增长一倍，预计到2020年人类拥有的数据总量将会达到惊人的35万亿GB。在这些新增的数据中，绝大部分是传统技术难以处理的非结构化数据，比如音视频、图片、网页等，这些数据大概占总数据的百分之九十。未来将是一个以PB（1024TB）为单位的，结构与非结构数据信息的新时代。

随着大数据时代的到来，计算机系统的数据分析和数据挖掘功能日渐强大，决策所依据的信息全面性越来越高，根据数据作决定的理性决策在迅速增多，而以往“拍脑袋”盲目决策的情况正在急剧减少。同时，由于云计算的兴起，人们得以高效率地驾驭海量数据，生产有价值的决策信息。在未来，基于大数据的决策将生成更多新奇有效的，解决重大问题的方案。伴随着大数据决策的发展，也许以前单纯依靠人类自身判断力的领域，最终都将被普遍改变甚至取代。

8.2 统计预测和决策概述

人们很早就开始利用统计方法对数据系统进行分析挖掘，进而预测和决策。统计预测与决策给各行业的发展走势提供重要的辅助引导，在此，笔者简单介绍统计预测与决策的基本概念和常见方法。

8.2.1 统计预测的作用及方法

统计预测可归属于预测方法的范畴，是利用科学的统计方法研究事物在未来发展变化的趋势及方向的预测方法，可以进行定量预测，同时计算概率置信区间。这种预测包括数学计算和直接判断。统计预测的方法论性质和统计学的方法论性质是相同的。

在市场经济条件下，预测的作用主要是通过每个企业或行业内部的决策以及行动计划来体现的。统计预测作用的大小往往取决于预测结果所能产生的效益的多少。影响预测作用的因素有多种，如预测费用的高低、预测方法的难易程度、预测结果的精确程度等。预测费用主要包括资料的收集、整理和使用费，计算费用，设计预测程序的费用和工作人员的劳务费用等。不难看出，预测费用的高低直接影响预测结果，而预测方法的难易程度又直接影响着预测费用的高低。至于预测结果，一般而言，精确度高的预测结果比精确度低的作用更大。虽然花费更多的费用、时间有可能得到更好的预测结果，但使用这部分额外的代价去取得额外的精确性是否值得，也是一个值得思考的问题。

对于统计预测方法来说，最基本的作用是将历史资料中并存的基本轨迹与误差分开，用以研究其形态的变化。通常是采取对资料进行拟合某种模型的方法把轨迹分离出来的，这种拟合模型要尽可能全面而精确地反映出有规律性的轨迹。误差又可称为残差或剩余项，呈随机性，残差的随机性研究是统计预测的一项重要内容。

统计预测的方法按照性质划分可大致分为三大类，定性预测法、回归预测法和时间序列法。

1. 定性预测

定性预测是一种以逻辑判断为主的预测方法。预测者依靠熟悉的业务知识、丰富的经验以及具有综合分析能力的其他专家，根据已掌握的直观材料和历史资料，以及个人的经验和分析判断能力，对事物的未来发展做出性质和程度上的判断，接着，再通过一定形式来综合各方面的意见，作为预测未来的主要依据。定性预测主要有两个特点：一为着重对事物发展的性质进行预测，主要依据人的经验以及分析能力；二为着重对事物发展的趋势、方向和重大转折点进行预测。

定性预测的优点在于注重事物发展在性质方面的预测，具有很大的灵活性，充分发挥人的主观能动性，并且简单迅速，节约时间和金钱。但是定性预测易受主观因素的影响，缺乏对事物发展作出数量上的精确描述。

2. 回归预测法

回归预测法是一种用来研究变量之间的相互关系的数理统计方法，应用回归分析从一个

或多个变量的值去预测因变量的值。回归预测运用样本数据确认其变量的相关性，再进行误差检验，最后运用模型进行预测分析。在这种预测方法中，因变量的预测值要由并进的自变量的值来推，故而该方法考虑了时间因素和变量间的因果关系。回归预测法一般可分为非线性回归预测法、一元线性回归预测法以及多元线性回归预测法等。

回归分析预测法的步骤可以归纳为以下五点。

（1）首先根据预测需求来明确自变量及因变量。

通常确定了预测的需求是什么之后，就可以明确因变量。比如预测的需求是下一年农作物的产量，那么产量Y就是因变量。而农作物品种、施肥量、气候等影响因变量（产量Y）的相关因素就称为自变量，同时还要从中选出具有主导影响的因素。

（2）接着建立回归预测的模型。

在自变量与因变量的历史统计资料计算基础上，建立一个回归分析方程，也就是回归分析预测模型。

（3）然后是相关分析。

回归分析是用数理统计分析来处理具有因果关系的自变量和因变量的。使回归分析预测模型有意义的前提是自变量和因变量之间的确存在着某种关系。所以，进行回归分析首要解决的问题是明确自变量的因素和因变量的预测对象有无相关性，相关程度怎样，以及有多大的把握判断它们之间的相关程度。通常都是先对变量进行相关分析，求出相关关系。相关程度则主要根据相关系数的大小来进行判断。

（4）再次计算预测误差，检验回归预测模型。

一般回归方程要进行各种检验，并且计算得到的误差要比较小，这样回归预测模型才能够用于实际预测，才可以将回归方程作为预测模型并进行预测。

（5）最后计算并确定预测值。

根据检验后确定的回归预测模型对预测值进行计算，并综合分析该预测值，然后确定最后的预测值。

另外，使用回归预测法的过程中还要注意一些问题，如变量之间应存在相关关系；现象之间的依存关系使用定性分析进行判断；需运用合适的数据资料等。

3. 时间序列法

利用按时间顺序排列的数据预测未来，是一种常用的方法。事物的发展变化趋势能延续至未来，常以时间序列的平稳性或准平稳性反映在随机过程理论中。准平稳性即经过某种数据处理（如一次或多次差分运算）后，时间序列呈平稳的性质。

多种因素影响着时间序列的变化，总的来说可以概括为四种变动因素：周期变动（C），以某一时间间隔为周期的周期性变动，诸如复苏与危机的交替；长期趋势（T），在整个预测期内事物所呈现出渐增或渐减的总倾向；季节变动（S），以一年为周期的周期变动，如旅游行业销售额的季节性波动；不规则变动因素（I）。

时间序列预测所需要的是序列本身的历史数据，因此这类方法应用得非常广泛，具体方法有时间序列分解法、指数平滑法、移动平均法、趋势外推法、自适应过滤法、灰色预测

法、平稳时间序列预测法、状态空间模型等。

由于现代预测方法的发展，使得各种方法往往交叉运用、相互渗透，很难做出明确地划分，因此，上述分类并不是绝对的。同时在选择预测方法时，还应当考虑适用性、经济性和精确性。根据具体的情况综合考虑，从而选择出最合适的预测方法。

8.2.2 统计决策的概述及方法

统计决策就是为了解决现实中出现的问题，实现某些特定的目标，根据客观的可能性，在充分搜集并全面分析了相关信息之后，提出解决问题和实现目标的各种可行性方案，根据评定标准和准则，选取合适的方案并实施。作为解决问题达到目标的方法和途径，决策具有未来性、选择性和实践性这三大特征。同时决策还是一项系统工程，其基本因素包含四个：决策主体、体现决策主体利益和愿望的决策目标、决策对象和决策环境。

完整的统计决策必须经历以下几个步骤。首先是确定决策目标，有效决策的前提是具有合理的目标，目标是决策活动的出发点和归宿，也是评价的依据。其次是拟订备选方案，可建立模型，要善于抓住关键因素和变量、参数及逻辑关系，最终选定参数和各种变量的数学公式，有时还需建立模糊模型及随机模型；若不方便建立模型，也可利用另外的数学分析方法帮助决策。再次是方案抉择，在分析各个可行方案时给出相应的分数评判，只需要按照选择的要求和标准，由决策者最终拍板确定即可。最后是方案实施，也是最重要的一环，确定方案后，组织人力、物力、财力等各种资源，实施决策方案。在实施过程中要注重监督和及时反馈信息，以便随时调整方案，做出更符合客观实际的决策方案。

统计决策方法可分为风险型决策方法、贝叶斯决策方法、不确定型决策方法以及多目标决策方法等。

风险型决策，面临的问题应该是明确的，然而在未来的决定因素中，对于可能出现的结果并不能充分肯定，在这种情况下，根据若干个可行方案执行后出现的不同结果和结果出现的概率作出决策。因此，这样的决策会有相应的风险，而决策者也要承担一定的风险。但是既然知道了结果和概率，就可以通过损益矩阵分析法和决策树法等来帮助决策者进行量化决策。

利用贝叶斯定理，求得后验概率，并据此进行的决策方法，称之为贝叶斯决策方法。

不确定型决策与风险型决策在条件和状态方面都比较相似，不同的是各种方案将来会出现哪一种结果这种概率无法预测。在思想方法上，不确定型决策靠他人的经验推断，或者靠主观判断，都具有一定程度的随意性。当然，不确定型决策也可以采用计算公式来帮助决策。

统计决策中的目标往往不止一个，例如企业目标决策中，企业不仅要追求经济目标，还要承担非经济目标，以及社会责任等。类似这样的企业目标决策问题均具有多目标的特点，可以把这类决策方法称为多目标决策方法。

在激烈的市场竞争中，科学的统计决策起着由目标到结果的中间媒介作用，发挥着避免盲目性和减少风险性的导向效应。

8.3　大数据预测决策的关键

预测是人们通过对客观事实历史和现状进行科学的调查和分析，由过去和现在去推测未来，由已知去推测未知，从而揭示客观事实未来发展的趋势和规律。预测作为一种手段，能为人们提供关于事物未来的信息，为决策者提供科学的决策依据。在决策全过程的每一个阶段都离不开预测，预测贯穿于决策的全过程。

人们能够预测的前提是事物的运动、变化和发展都呈现出一定的规律性。然而，事物的规律通常以隐蔽的形式存在，受创造性思维能力、知识经验、对历史资料的掌握等多方面因素的影响。要正确认识到事物潜在的规律，作出准确的预测决策，向来是决策者的难题。现今，大数据技术的发展，使预测决策更为客观、科学。

海量的数据，无论是结构化、半结构化或是非结构化的数据，都使人们对行业有了可能全面分析预测的基础。决策者可以通过对这些数据进行系统、全面地分析挖掘，从而预测未来某些事件的发生概率、走势，并根据预测结果辅助自己作出重大的决策。对大数据的分析、预测大大降低了管理者的主观判断风险，客观的数据使得做出的决策更具科学性。

云计算、大数据和社会化媒体等技术提供了大数据决策的技术手段。在大数据时代，如何利用大数据进行决策已成为企业、机构、政府部门等的工作重点。

1. 手握大数据源

欧洲消费者委员会委员梅格莱纳·库尼瓦说：“数据是一种新型石油”。在大数据时代，数据渗透到各行各业之中，已成为企业的创新驱动力与核心资产。企业的核心竞争力将取决于拥有数据的质量、规模和收集、处理、分析、运用数据的能力。数据越多，相关度和质量越高，找出有用信息和得出结论的概率就越大。而那些占有“大数据”资源的先天优势群体，无疑在有效利用好数据，打破现有的传统格局上更具有优势。掌控数据就能够支配市场，也意味着高额的投资回报。据有关数据统计显示：在美国，每提高10%的数据智能化，服务以及产品质量将提升14.6%。但是，在大数据时代进行智能化决策分析，企业首先要拥有大数据源。数据来源除了企业的ERP、HCM、CRM、OA等系统内部的数据外，同行资料甚至应该包含不同行业的电商、社交、宏观经济、上下游、互联网、物联网等外部大数据。事实证明，不同行业的数据也可能对自己所在的行业存在影响。如Twitter网根据Twitter用户的集体情绪，就能够预测股市的涨跌。故而要实现智能化决策，首先要尽可能通过云平台实现数据大集中，形成企业数据资产，为大数据决策做准备。

2. 培养大数据科学家

互联网上，每一天，百度要处理的搜索请求大约为数十亿次，联通的用户一天上网产生的记录就能达到10 TB，淘宝网站会产生数千万笔交易，新浪微博的用户会产生超过1亿条的发博量等并且数据已从结构化数据转向非结构化、半结构化和结构化混合型数据。当拥有海量的数据之后，如何进行数据挖掘，将数据转换为知识，再将知识付诸于行动，这成为大数据科学家存在的意义和必然结果。大数据科学家从大量的、不完全的、有噪声的、模糊的、随机的实际数据中，提取隐含在其中的、人们事先不知道但又是潜在有用的信息和知识。经提炼

分析后采用先进的技术进行可视化展示，使得决策者能够清晰地快速洞悉数据背后隐藏的商业价值，从而运筹帷幄，决胜千里。因此，组建包括大数据科学家、大数据工程师、大数据分析师、商业情报分析师以及事业部用户在内的大数据科学家团队对于大数据决策势在必行。

3. 开发大数据预测分析软件

在培养顶尖的大数据科学家的同时，还必须不断开发适应时代需要的更新、更好的软件，因为大数据科学家在预测分析时，要借助预测分析软件来评估分析模型和规则。预测分析软件通过整合统计分析和机器学习算法来发挥作用。目前，IBM SPSS和SAS是两个数据科学家常用的分析软件，R项目则是一个非常流行的开源工具。如果数据量大到“大数据”的程度，那么可能还需要一些专门的大数据处理平台如（Hadoop）或数据库分析机（如Oracle Exadata）等。据预测，未来百分之九十的数据是传统技术难以处理的非结构化数据，诸如音频、视频、图片、网页等，因而需要开发更多适用于处理半结构化数据和非结构化数据的软件。

对大数据的预测决策给企业决策管理带来了极大的惊喜，在这信息爆发的时代，谁能领先掌握大数据技术，谁就有可能在竞争中领先一大步。成功的企业除了要有好的决策管理方法外，企业本身也应该具备良好的原则。

诺贝尔奖获得者赫伯特·西蒙（Herbert·A·Simon）说“管理就是决策”。每个企业都存在决策，决策是决定管理工作成败的关键。在大数据时代，如何成为决策领先的企业呢?我们有以下建议。

首先，决策领先企业是系统化和量化的。

企业系统化和量化不仅仅是数学计算和使用自动化来更快、更好地作出决策，它也与管理企业决策的更好收益有关。决策收益应该体现在这样几个方面，决策的准确性，即企业作出盈利的、有针对性的决策；灵活性，公司可以在繁重的工作中作出决策；一致性，企业在不同的渠道、业务部门和地区以相同的方式作出决策；成本，公司能够实施自动决策，减少工作步骤和降低运营成本；速度，公司可以实时作出决策，加快业务流程。唯有实现系统和量化的决策管理，企业才能始终保持领先地位。

其次，决策领先企业是不断学习和提高的。

企业应用大数据决策，除了企业领导必须有与时俱进的观念，能够充分认识到应用大数据决策的好处，还应该建设以大数据决策为基础的顶层设计，打造企业为学习型组织，并建立制度、平台和流程，形成知识不断成长的企业文化。或者与外部数据公司合作，比如百度已初步形成一个海量知识数据汇集、互动与共享体系，为知识提供、搜索、分享、利用创造了高效率平台，通过海量离散知识提供者和海量知识需求者相匹配，改变了学习活动的时空限制，使企业在培训中不断学习和提高。

最后，决策领先企业是大胆且富有创意的。

全球最大的家用电器和电子产品零售集团百思买提出了利用客户细分来改变经营形式，这被认为是一个革命性创举。实际上，百思买的高层管理人员在看到客户细分后的客户体验差异之前，他们几乎没有考虑过要改变经营的形式，在所有的渠道提供个性化体验。加拿大最大的零售连锁企业加拿大泰尔开发了信贷风险预测模型，该模型能根据信用卡客户的购买

信息，对客户进行高风险和低风险的区分。事实证明，加拿大泰尔运用对客户群的分类作出了更好的运营决策。这些案例告诉我们大胆和创意是决策领先企业至关重要的因素。

在大数据爆发的时代，各行各业唯有善于利用数据资源，顺应时代发展的趋势，掌握大数据决策的关键，才能在大数据的浪潮中处于领先，立于不败之地。

8.4　大数据分析用于商业的预测决策

现今铺天盖地的大数据开始充斥在各行各业中，无论是医疗器械、零售业、物流业还是电力通信、金融服务，几乎每个行业领域，伴随着移动互联网、移动终端和数据感应器的出现，所产生的数据都以超出人们想象的速度在迅速增长，伴随而生的大数据技术也开始飞速发展。与此同时大数据分析用于商业预测决策的成功案例也越来越多。

8.4.1　乐购——分析客户消费信息

乐购是英国领先的零售商，并跻身于全球三大零售企业之一，就营业收入而言，乐购是网上最大的百货服务商。

乐购的成功之处在于早在20世纪90年代中期，公司就累计储存了超过一百万老客户的数据，并利用分析工具了解将这些数据作为和服务于个性化的需要。在1994年，乐购雇用了一家小型分析公司Dunnhumby，Dunnhumby帮助乐购从收集到的客户数据中挖掘出了更多的商业价值。而这些收集的数据主要是从会员卡中获得。乐购会员卡追踪并保存每个成员的每一笔交易，随着每笔交易的完成，客户和相关购物的详细信息不断增长。从获取的数据中，乐购了解到客户最常去的乐购店的地方，客户买了什么，客户多久逛一次乐购，客户的年龄段和生活方式等。通过开辟网络市场，乐购服务的普及和数据收集分析有了质的飞跃发展。乐购的电子商务团队制作了一个“最爱列表”系统，能根据客户的购买记录向客户展示他们最喜爱的商品，从而省却了浏览几千种商品目录的时间。乐购使用分析学精确描述了顾客的购物模式，调整客户购物体验，同时提高了公司的盈利能力。

现在，乐购收集大数据客户信息用于分析。《经济学家》报道认为，乐购的会员卡项目从130万会员购买55000种产品的行为中，产生了令人无法想象的庞大数据。根据大数据信息进行客户分析、客户细分，建立预测模型，预测客户对产品的需求量和未来的销售情况。

8.4.2　Netflix——了解客户的真正需求

Netflix是美国一家在线影片租赁公司，提供互联网随选流媒体播放、定额制DVD、蓝光光盘在线出租业务。Netflix将百视达公司作为主要竞争对手。百视达是一家曾击败许多租赁店的视频连锁店，其分店遍布美国各地。

Netflix一开始就借助于互联网，使客户群快速增长。客户在互联网上完成注册、下订单、通过邮政服务派发DVD。同时Netflix公司利用预测分析技术，了解客户希望购买或租赁什么电影和电视节目。通过数学技巧和商业敏感性持续改善客户体验感。最终Netflix通过使用它的推荐引擎——Cinematch超过了百视达公司。这款引擎运行专有的预测算法，分析客户

关于电影的购买模式和评价等级，预测客户的喜好并优化库存状况。Netflix回馈给老客户的是给予更多他们想要的东西和选择。Netflix利用客户每一次购买获得的信息，建立了一个专有的档案。Netflix的成功在于发现客户需求，了解客户希望的获得方式，满足客户的需求。

2013年，Netflix在美国已拥有2700万订阅客户，每天客户在Netflix上产生3000多万个行为，同时客户每天还会给出400多万个评分以及300万次搜索请求。Netflix的工程师借助网上这些数据进行分析后，发现喜欢BBC剧、导演大卫·芬奇和老戏骨凯文·史派西的客户存在交集，为此Netflix决定投巨资买下BBC电视剧《纸牌屋》的版权，并请来大卫·芬奇担任导演，凯文·史派西担当男主角。《纸牌屋》成为了Netflix网站上有史以来观看量最高的剧集，并红遍美国及40多个国家。《纸牌屋》开创了大数据应用在电视剧制作的先河，这是大数据帮助人们做出前瞻性决策的实例之一。

8.4.3 哈拉斯——使用客户数据

哈拉斯，世界上最大的博彩娱乐集团。哈拉斯的首席运营官加里·洛夫曼是哈拉斯首屈一指的分析师，他分析后发现，最好的客户并不是“拥有黄金袖扣，豪华轿车的豪赌者”，而是中年的中产阶级，他们喜欢玩老虎机。这些客户只占总体客户的26%，然而他们却贡献了82%的公司收入。很明显，如果能够增加这个客户群，公司的收入将会飙升。于是哈拉斯推出了“Total Gold”客户忠诚度项目，在客户玩老虎机时会获得相应的积分，根据积分可以兑换各种礼品、礼券。

如今，哈拉斯的客户分析已成为其商业模式的核心，它建立了一个客户洞察团队，专门进行管理和挖掘客户数据，哈拉斯的客户分析能力在同行业中是无与伦比的。哈拉斯有一个300GB的交易数据库，是通过追踪数以百万计的个人交易而获得的。信息系统为项目的基础，收集了大量的客户数据。再把数据基于行为进行细分，将数据按照行为和收入划分，从中寻找挥金如土的人。通过捕获来自多个地点的个人及群体行为和消费模式的数据信息，预测客户的行为，进而进行各种营销决策，针对什么样的客户采取什么样的渠道，在不同的渠道发布不同的消息。例如当一个月均消费1000美元的顾客，在几个月里没有再踏足哈拉斯时，数据库触发器就会发信或打电话邀请他回来。

哈拉斯运用分析技术对客户大数据信息进行挖掘，从而帮助它预测决策，使公司成为全球博彩业的巨头。

8.4.4 大通银行——决策树方法分析按揭数据

大通银行成立于1799年，目前在65个国家有办事机构，总资产超过3960亿美元，是世界大型商业银行之一，对美国现代金融和经济有着巨大的影响。

早在20世纪90年代中期，大通银行发现他们在按揭贷款评估中总会出现很高的失误率。为此，大通银行请来了一位商业科学家丹·斯坦伯格，希望借助斯坦伯格研发的系统来评估、处理大量的银行按揭，同时希望能够用预测值来判断个人按揭贷款的未来价值，由此来确定是否将这些贷款转让给其他银行。

斯坦伯格根据要求组建了一个小规模的专家团队，将分类回归决策树方法用于分析大通

银行的按揭数据。在这个按揭贷款预测评估程序中，斯坦伯格所要预测的对象是在未来三个月内将提前还款的按揭贷款人，而所要采取的行动则是评估按揭贷款的价值，决定是否将这笔贷款转让给其他银行。最终该预测系统成功地帮助大通银行预测了数百万份按揭的风险。

大通银行从该预测项目中受益匪浅，并于2000年购买了JP摩根组建摩根大通集团，目前按资产总额测算，摩根大通已经是美国市场上最大的金融机构。

8.4.5　好事达——采用高级预测分析技术

好事达保险公司（Allstate）每年根据投保车辆的状况来预测如果出现交通事故时车内人员最有可能的受伤情况，以此为根据来调整保险方案，这项预测每年为公司节省了将近4000万美元。目前已有不少保险公司在精算中采取高级预测技术，大大提高了公司的盈利。

上述成功案例无一不揭示着预测决策的重要性。成功的预测决策给企业带来的是无可估量的价值。海量的数据使得预测分析更为依赖于客观的数据而不是某人的主观判断。在过去，人们下决定时往往是根据自己的经验判断，有时甚至是个人的心情喜好。当这些判断者处于高位时，他们的判断对企业的存亡则尤为关键，领导者一个正确的判断很有可能使公司抢占先机，在激烈的竞争中脱颖而出，从此迈向新的里程碑；而一次错误的决策则很有可能使公司错失良机，甚至一蹶不振。故而在传统的企业中，一位经验丰富、目光敏锐的领导人非常重要。世界著名咨询公司美国兰德公司的一项经典调查认为，世界上每1000家破产倒闭的大企业中，85%是因为企业管理者的决策不慎造成的。可见，办好企业，关键在于领导的决策水平。然而，个人的经验、人生阅历是有限的，且受个人的性格、偏好等因素的影响，要想长期做出正确的判断是一件极其艰巨的，甚至难以实现的任务。大数据预测分析的诞生，可帮助人们解决这一难题。

大数据预测决策是一门极其重要的学科，目前正在不断成长中。随着预测分析技术的逐步完善，预测决策带给我们的将是无法估量的价值。

8.5　大数据时代给政府决策管理带来的机遇与挑战

大数据是继云计算、物联网之后IT产业又一次颠覆性的技术变革。“大数据”时代的来临，对政府决策管理机遇与挑战并存。怎样应对大数据，运用大数据，并主动顺应大数据时代来改进政府决策管理，是现阶段政府部门面临的重大课题。

8.5.1　大数据提升政府的决策管理能力

在当今世界，大数据分析技术已为世界多个国家所重视和运用，政府部门越来越注重运用技术手段对数据资源进行深度的价值挖掘，以满足日益增长的精细化、科学化管理的需要。大数据分析技术已成为政府施政的主要工具之一。甚至发达国家都在运用大数据精准营销来提升政府机关、个人的商业价值和形象。奥巴马就是其中的佼佼者，在2012年的大选中，奥巴马有效地利用了社会化的精准营销，获得了大胜。

“大数据”时代已经到来，2011年3月11日，在日本大地震发生后仅仅9分钟，美国国家

海洋和大气管理局（NOAA）就发布了详细的海啸预警。其快速反应归功于世界范围内庞大的海洋传感器网络。根据海洋传感器收集的实时数据，NOAA进行计算机模拟，并制定出详细的应急方案。这一机构花费巨大，每年的预算都高达10亿美元，美国政府却乐此不疲，因为数据关乎生命。

对政府部门而言，大数据所能带来的巨大能量已经显现，甚至已经超过了技术改进产生的效益。与互联网的发明一样，大数据分析绝不仅仅是信息技术领域的革命，更是建设数据政府，引领政府智能决策的利器。一般来讲，大数据主要在以下几方面对政府决策管理起作用。

在政策制定阶段，数据分析是决定政策质量高低的关键性因素。通过对历史数据的有效分析，可以吸取教训，总结经验，为新计划的制定提供宝贵的借鉴。对当前及未来影响政府活动的可能因素进行量化分析，辅之以同期其他国家（地区）同类活动的运行比较，可以为政策的制定提供更为直接、更加重要的参考。如广西采用了PADIS系统（国家人口管理与决策信息系统），通过输入本地的相关具体数据和备选政策，最终作出修改自治区人口和计划生育条例的决定，条例生育两个子女的条件从以前的“夫妻双方均是独生子女”放宽为“夫妻一方是独生子女”。

在政策实施阶段，大数据分析技术可以有效地监控政策实施的情况。首先可以通过数据分析监控，及时掌握政策是否按计划如期实施，了解影响政策的顺利实施的因素有哪些。此外，对于政策实施过程中出现的一些问题或失误，数据分析技术可以快速、准确地反映给决策者，从而能在第一时间提出补救或修正措施。

在政策评估阶段，数据分析同样有着不可忽视的作用。如政策的实施是否发挥预期的作用，实施后又产生了哪些其他方面的后果等。这些问题都需要通过科学的数据分析来解答，同时也对未来政策的制订有着极其重要的借鉴意义。

随着社会经济文化的发展和进步，公众对政府和职能部门的要求也越来越高，集中表现为要求提高行政效率和透明度、创新工作方式、提高对社会的服务能力等。大数据技术的发展给政府决策管理能力的提高奠定了基础。如中国的耕地和粮食生产管理问题，以往依靠镇、县、市、省再到国家统计局，统计人员层层上报，费心、费人、费力，但由于基层出于地方利益的考虑，且受工作能力、精力所限，其数据上报的准确性令人怀疑。而如今政府采用大数据建模的方法，通过遥感卫星来识别图像，把中国所有的耕地标识、计算出来，然后把中国的耕地网格化，对每个网格的耕地抽样进行跟踪、调查和统计，接着按照统计学的原理，计算出中国整体的耕地、粮食数据。这种采用大数据建模的方法使政府更易获得真实的决策数据。此外，大数据使得政府真正拥有执政为民的能力，维护人民的权益。例如，利用数据对广大消费者的质量诉求进行统计归纳，就可以迅速知晓哪些公司、哪个地区、哪类产品质量问题比较多，从而更有针对性地进行整治管理和监管，才能对百姓提供更好的服务。同时，可利用大数据技术挖掘海量的市场商品数据，如商场退货信息、消费者投诉信息、医院伤患医疗信息、当地政府执法和监管信息、公检法机关执法信息、媒体传播信息、网络交流信息、科技研究信息等，通过对这些数据进行全面地挖掘、整理，能够找出一些产品的质量安全问题或潜在的质量风险，然后采用简单的技术进行检测验证，这样政府管理部门就能提前了解风险，降低风险，化解风险，防患于未然。

通过大数据，传统的管理方式将被颠覆，与此同时自身管理的有效性也得到提高。以往，政府部门内部的管理方式主要依赖于层级组织以及严格的流程，依赖信息的层层汇集来制定正确的决策。再通过决策在组织中传递和分解，以及规范流程，从而确保决策得以贯彻实施。在传统各方面技术有限的条件下，这无疑是一种有用的方法，然而这种方法显得费时而笨拙。如今在大数据时代，可以重构质监部门的管理方式。通过实施远程数据传输、监控，能够对企业实行实时数据比对监管、对新的监管数据实时录入，可以进行远程在线指挥。通过大数据的分析和挖掘，政府不需要再依靠庞大的组织以及复杂的流程，这样可以改变以往的层层报告，使大量的管理业务能够根据既定的规则进行自主决策，节省了时间。数据化地管理工作人员的行程、任务，通过定期统计分析，可以及时准确地了解工作人员的工作量、完成任务情况、工作绩效等。这样通过实实在在的数据来说话，改变过去靠印象管理，提拔人才的方式。

总而言之，在这样一个信息爆炸、技术发达的新时代，大数据是一种可再生的重要资源。该资源能够被国家和各行业反复利用，是一种越挖越多、越挖越值钱的资源。2012年，联合国在发布的大数据政务白皮书中指出，大数据对于联合国和各国政府来说是一个历史性的机遇，人们如今可以使用极其丰富的数据资源，来对社会、经济进行史无前例的实时分析，帮助政府更好地响应社会和经济运行。因此，要尽快适应新形势，加强新知识的学习，及时了解大数据、运用大数据，通过云计算等新技术的运用，促进政府智能决策管理工作不断走上新的台阶。

8.5.2　大数据浪潮中政府面临的挑战

大数据给政府带来了一个全新的、历史性的机遇，其背后隐藏的价值更令人激动，然而大数据时代来临同样使政府决策管理面临了巨大的困难和新挑战。

1. 政府的信息化建设水平

大数据对政府的信息化建设水平提出了挑战。相对于过去数据处理方式，无论是过程（从数据搜集到数据处理），数据类型（结构化、半结构化和非结构化），处理标准（标准统一与标准各异），还是处理对象（只面对样本或面对庞大的总体），都存在着巨大的差异，尤其是数据的搜集和处理难度也明显更大。大数据中包含更多的是非结构化数据，如图片、视频、文字，处理这些数据有着极其重要的意义。在美国就有一例通过视频数据分析解决的爆恐案。美国波士顿在各个地方都装有高清摄像头，某天，一个可疑的黑色背包突然出现在马路边上，根据这个异常信号，视频中心立刻就报警了。爆炸事件发生后，回溯到这段视频，根据视频提供的信息，警方快速地锁定了做这个行为的嫌疑人，接着连续性跟踪，很快把犯罪嫌疑人绳之于法。侦破这个爆炸案件就涉及到大数据技术，如何在海量的高清视频中提取这种异常节点的数据，以及如何在多个非连续的摄像头中间，找出某人相关的行动轨迹，这在以往的传统技术里难度是很大的，需要使用新的技术处理方法。因此，如何将这些非结构化数据进行结构化处理，对海量的数据进行分析挖掘，培养更多的大数据科学家是政府信息化建设中要面对的一个重大课题。

2. 如何去噪音留信号

大数据指的是所涉及的数据量规模巨大、关系混杂、动态持续、变化不定，需要用先进的技术和工具，在合理时间内实现数据的撷取、存储、分配、提炼、处理、集成和分析，并从中挖掘出有价值的信息。然而如何才能快速而准确地分析挖掘数据，并根据预测进行决策，这是大数据预测决策的重点、难点。目前大数据公认的四个特征里面，价值密度低一直是我们面临的难点。在海量的数据里面，大部分数据都只是噪音而已，真正有用的信息非常少。就目前而言，分析和挖掘数据的增长能力远远跟不上实际需求，太多的数据需要发掘。在信息爆炸的大数据时代，政府所要做出的预测的速度和数量都在不断增加，然而现实世界中的很多预测都失败了，也因此而付出了巨大的社会代价。例如2008年金融危机、卡特里娜飓风以及禽流感肆虐等，由于预测者没办法从杂多的噪音中把信号挖掘出来，忽视了真正的信号，最终导致错误的预测，从而付出了巨大的代价。因此政府决策管理面临的一大难点就是如何提高分析挖掘数据的能力，将正确的信号从混杂了噪音的数据中提取出来，从而为正确预测决策打下良好的基础。

3. 正确地运用大数据工具

大数据预测决策是顺应时代发展的需要而诞生的，将给政府决策管理带来一场新的改革。然而在这巨大的喜悦面前，我们更应该时刻保持清醒，不要盲目地被大数据“牵着鼻子走”。要知道，人才是信息社会的主体，我们应当积极发挥主观能动性，在对数据进行分析挖掘并得到有用的信息时，要善于将信息与丰富的经验相结合，让大数据预测辅助政府决策，而不是大数据预测来决定政府决策。如2008年谷歌推出了流感趋势系统来监测全美的网络搜索，通过寻找与流感相关的词语，比如“咳嗽”和“发烧”等，谷歌根据这些搜索来提前预测流感的爆发，并在2009年成功地预测了甲型H1N1流感的爆发，然而2013年谷歌预测的流感信息与美国疾病控制中心报道的结果相差甚远。有专家认为“大数据傲慢”（Big Data Hubris）和算法变化是导致谷歌预测发生错误的主要原因。“大数据傲慢”即人们认为大数据可以完全取代传统的数据收集方法，而非作为后者的补充。在大数据的时代，政府应清楚大数据技术的定位，善于运用大数据技术工具来帮助政府决策。如何正确地运用大数据工具而不是盲目地依赖于大数据，这也是政府目前的一个难点。

4. 保护信息隐私权

数据的价值、力量和意义让数据变得很敏感，而最令人关注的是数据的收集和预测会不会侵犯了个人的隐私。的确，在大数据时代，人人几乎无处遁形，拿互联网领域来讲，顺着社交网络的这一张大网，总会找到人们的蛛丝马迹。人们通过电脑、手机等电子设备在网络上进行的每一个操作，都被服务器记录了下来，而且这些被记录的个人信息将面临着二次利用。例如目标超市根据消费者采购商品的种类等来分析怀有身孕的消费者，从而扩大营销受众群体；药店根据人们的医疗信息推测其身体健康状况，是否患有某种疾病，从而向目标客户推销药物……这种种的情况会让人们感觉自己的隐私毫无遮掩地暴露在他人面前，生活中无处不在的“第三只眼”更是引起了人们的惊恐。因此，政府应当保护公民的隐私权，对数据的隐私权进行保护。进行数据搜集以及分析的人员也都应当具有保护数据隐私的职业道

德和高度的责任感。在发掘大数据价值的过程当中，为了实现价值的准确匹配，有时会无法避免地利用到这些隐私数据。在这种情况之下，必须通过法律手段来确保个人隐私数据的权利。就公民个人而言，在享受大数据时代所带来个性化服务的同时，也要加强风险防范意识，在有可能留下隐私数据的情形下，能够充分考虑到因为隐私暴露而有可能给自己带来的不利，同时采取相应的防范措施来保护隐私。而政府方面则应该界定信息隐私权，即使概念的界定会带来争议，但这些可以在广泛的讨论上达成共识。我们应当在这个基础上，对数字隐私权进行基础设施建设。与此同时，推动相关立法进程来打造一个良性的信息生态环境。

5. 数据安全管理

大数据意味着大量的数据，同时也意味着更加繁杂多变的数据，毫无疑问，大数据的价值将吸引更多的攻击者。在网络空间，大数据往往更容易被黑客“盯上”。数据的大量汇集更便于黑客攻击，黑客成功攻击一次就可以获得巨量的数据，这无形等于降低了黑客的进攻成本，提高了攻击者的效率。此外，大数据存储带来了新的安全问题，威胁着现有的存储和安防措施。数据大集中的后果是复杂多样的数据存储在一起，很可能会出现数据存放不合理的情况。例如企业将某些生产数据放到经营数据存储的位置，致使企业安全管理不合规。政府也同样会面临安全管理不合规等情况。此外，大数据的大小也与安全控制措施能否正确运行息息相关。大数据技术不仅能给政府的安全防护手段的更新升级带来帮助，也给黑客的攻击加强了手段。可以想象一下，如果安全防护手段的更新速度无法跟上数据量非线性增长的步伐，那么大数据安全防护的漏洞就会暴露。最终，大数据技术将成为黑客用来对付政府的手段。与此同时，大数据技术也给黑客发起攻击带来了更多的机会。黑客利用大数据进行僵尸网络攻击，可以同时控制上百万台傀儡机并发起攻击。如韩国就曾遭遇过大规模APT攻击事件，在2013年3月，韩国多家大型银行和媒体相继出现电脑丢失画面的情况，系统几乎瘫痪。根据国家互联网应急中心统计显示，2012年移动互联网恶意程序数量骤增至162981个，较2011年增长了25倍。这些事件敲响了数据安全的警钟。鉴于大数据的重要价值，政府更应加强数据安全的管理，把数据牢牢地掌握在自己手中。

8.5.3　政府以变革来顺应大数据时代

政府在大数据占有方面具有天然的优势，是大数据的最大受益者，因而也在建设大数据基础设施、培育大数据产业、培养大数据人才、完善相关标准和立法等方面负有更至关重要的责任。我国是社会主义国家，政府在资源配置方面发挥着极其重要的作用，其强大影响力是带动大数据加速发展的优势所在。但是，我国政府在大数据应用方面才刚刚起步，要利用好大数据，所面临的困难不仅仅是技术上的因素，更面临一系列的大转型、大变革。

对政府公共管理水平的变革，主要体现在以下5个方面。

（1）实现信息透明和共享。都说大数据是新型的石油，这宝贵的资源不仅能提高政府内部利益相关者（比如政府雇员和政府机构）的工作效率，它的再生性及其潜在价值也可以给公民、企业等带来直接或间接的价值，产生积极的经济社会综合效益。例如2006年起，中国人民银行上海总部公开了金融信息，从而催生了一大批金融信息咨询服务公司。其中上市的包括上海联和金融信息服务有限公司等5家公司，这些金融公司解决了十几万人的就业问

题。所以，实现信息透明和共享对政府和企业都是互惠互利的。

（2）通过人口细分和定制政策，有针对性地对特定群体进行公共服务，提高公众满意度和工作效率，减少不必要的开支。在以往的公共管理中，公共部门更倾向于为所有公民提供相同的服务。事实上，不同的人群往往有不同的需求。部门由于不能有的放矢，通常会在不必要的方面浪费巨资，同时也达不到有效服务公众的目的。德国联邦劳工局就采用数据分析的手段，对失业人员的失业情况、干预手段以及重新就业等做了研究，有效地区分不同类别的失业群体，从而有针对性地进行干预，提高了公共服务的有效性。据了解，该做法大大缩短了失业人员再就业的时间，改善了他们的求职体验，并使该局每年因此减少了100亿欧元的相关支出。

（3）为了增强内部竞争，进行公共部门的绩效评估，激励人员的工作表现，提高公共建设的效率，合理降低政府的管理成本，提升行政服务的质量。譬如，荷兰政府推行利用大数据来提升防洪能力。在数字三角洲的工程实施过程中，通过协调国家研究所、税务部门和环境部三方的财力、物力以及人力，研究如何采用大数据预测，改变防洪策略和整个荷兰水资源系统的管理工作。并通过IT技术以及更好的方式来处理可用数据，从而达到降低成本的目的。结合相应的数据推测发现，这项工程预计可以节省高达15%的荷兰年度水资源管理预算。

（4）利用政务智能代替或辅助人工决策，人的精力是有限的，要在纷繁复杂的数据中找出错误和虚假的信息，需要耗费巨大的人力物力，而大数据技术的发展为降低出错成本和防止福利管理中的诈骗等提供了可能。美国邮政署的计算机系统就可以自动扫描邮件的相关数据，通过与数据库中约4000亿条数据（如存放位置、路线、重量和体积等信息）的比较，筛选出“邮资欺诈”的邮件。每封邮件仅需要花50 ~ 100毫秒扫描。当检测出了异常，诸如邮资不足抑或邮票重复使用等情况时，系统就会对该信件实行拦截。由于该项目的实施，美国的邮资欺诈行为大幅减少，对“邮资欺诈”行为起到了很好的威慑效应。

（5）引导公共部门内、外部的创新。比如，商业、非营利组织、第三方等可通过开发大数据工具和分析来反馈给公共服务，帮助改善现有的方案，进而为公共部门创造新的价值。作为大数据惠民的一项重要探索，北京市于2012年10月推出了政府数据资源网的测试版，并面向企业及个人征集APP（应用程序）。由社会力量开发的“上下班路况”“交通英雄”“游北京”和“爱健康”程序已经可以下载试用。如“上下班路况”可对北京市全市路况概览、周边路况查询、交通执法站查询、上下班旅行时间预测等。这些程序给用户带来了极大的方便。

此外，大数据技术还给政府互联网管理、政府统计、应急管理系统等各方面都带来了极大的改革。

如在互联网管理方面，政府与北京国双科技有限公司合作成立了国家信息中心网络政府研究中心。国信中心网络政府研究中心副主任于施洋在国双数据中心启动仪式中表示，大数据对于政府最核心的价值就是建设开放的政府，一个有责任的智慧政府。在大数据的时代下，政府在信息化领域首先应该转变工作的观念，坚持改革开放，不搞封闭。同时政府在信息化领域中更应该发挥引导的作用，美国的大量案例证明，政府在大数据中，会带动整个大数据产业飞速的发展，这已经是一个事实。所以，政府在这个领域更应当起到表率的作用。另外，政府用信息化提升创新的能力是大势所趋。未来，大数据精准营销还给政府互联网管理带来更大的改革，政府互联网可以“精准地感知”和“有效地反馈”，更了解公众的需

要，及时服务公众，使政府未来在互联网和公众之间形成和睦和谐的互动，创造一个更美好的互联网环境。

政府是大数据时代的先锋，更应该把握好在大数据浪潮中前进的方向，顺应时代的要求和发展趋势，不断地改进创新，争取早日建立世界一流的智慧政府。

8.6　大数据时代的跨界与颠覆

互联网、移动通信和大数据对传统行业的重构已经成为中国经济新一轮快速发展的关键推动力。随着新生代的消费需求的变化，对产业结构和运营效率的要求越来越高，“跨界”现象正在重塑着传统产业格局。

8.6.1　大数据时代，颠覆浪潮席卷传统产业

大数据、云计算、移动互联网、搜索引擎、社交网络、物联网等新兴信息技术的发展，改变了传统的信息产生、传播、加工利用的方式，尤其是互联网、移动互联网和大数据有望成为中国经济新一轮快速发展的关键推动力。在此信息技术基础上越来越多的基于云服务的软件应用开始涌现，而在软件应用的背后则聚集了海量的数据。最终数据又需要通过电子产品来展现。从本质上，互联网与大数据是对传统产业的价值核心要素的分配，是生产关系的重构。大数据颠覆传统产业，提升运营效率和改变产业结构。21世纪是颠覆浪潮席卷传统产业的大数据时代。

大数据分析技术可以帮助人们发现规律和作出预测决策，将对经济、社会产生巨大的影响。而大数据时代首先给各行业带来的是颠覆传统的改变，从管理、构架、实施方式等方面，新的技术都使之焕然一新。其中受大数据影响最大的行业之一就是医疗行业。

新兴的医疗技术和设备正在改变医生的行医方式、临床评估方式和与患者交互的方式。在国外已经开始实现数字化寻医问诊。如美国在线医疗服务使医生或联合医疗服务提供者实现了实时、安全的电子化视频问诊。苹果、谷歌、百度等纷纷研制、推出了智能手表、智能手环，这些产品能随时随刻采集用户的体温、睡眠质量、血压、脉搏等数据，且这些数据可通过智能终端传递到远端，对用户的健康状态进行良好的管理。Augmedi运用谷歌眼镜技术，为临床和门诊大夫提供服务，通过自动化生成患者就诊文档，解决了难点问题，具有巨大的价值。大数据技术使得医生可以跨市、跨省甚至跨国学习了解其他顶尖医生的治疗经验，汇集、提炼全球优秀医者的治疗方案，进一步提升自己的医学专业和实战经验。同时，大数据技术让医学诊断在未来有可能逐渐演化为全人、全程的信息跟踪、预测预防和个性化治疗。著名的苹果公司联合创始人史蒂夫·乔布斯，曾经花费10万美元将他的癌肿瘤和正常的DNA进行测序，以期找到针对性的治疗方案。相信随着科学的进步，根据患者的遗传基因和临床信息的全民个性化治疗将不再是构想。

发现与预测是大数据应用的核心，交通运输行业利用这个特点使其服务水平得到极其大的提升和改变。大数据的虚拟性以及信息集成优势和组合效率较大程度上改变了传统公共交通管理的路径。大数据可以跨越行政区域的限制，有利于信息的跨越区域管理，利用大数据技术进行集成、检索、分析；挖掘出有价值的相关信息，尽可能地满足各种交通需求，从

而使实时交通障碍得以顺利解决。例如，根据全国高速公路收费数据，通过结合交通流量调查信息与重点营运车辆联网联控信息，可以比较准确地获知在某一时刻、某一区域的人员流量、车流量，甚至是物流信息（车联网），进而对未来几十分钟甚至更长时间内的路网交通状况进行预测。当发生突发事件时，能够实时判定对区域交通的影响趋势，从而及时采取相应的措施。此外，大数据还可以帮助警方破获车辆盗窃案。当车辆被盗后，数据中心能够调取整个城市内的摄像头监控视频，进行数据分析，得到被盗车辆的具体位置及时间信息，即通过对该车辆某一时刻在什么位置，及该车的行驶轨迹和城市道路运行情况等数据进行综合分析，进而推测出该车下一刻可能在哪个地方出现。根据预测结果，警方在预测点守株待兔即可。另外，大数据还能够检测套牌车，运用数据分析找出同样的牌号同一时刻在不同地方运行的车辆，使得套牌车无所遁形[①]。事实证明，先进的大数据技术，能帮助政府更好地配置公共交通信息资源。IBM研究中心针对交通的大数据管理这一主题，与加利福尼亚州运输部以及加利福尼亚大学伯克利分校的加州创新运输中心（CCIT）进行合作，目的在于预测上班族的交通条件。现在大数据在交通管理方面的作用日渐显著，一些大企业，如谷歌、苹果等都已利用大数据来解决交通问题。

大数据带给传统行业的除了管理、构架、方式等颠覆性的改变外，更让人震撼的莫过于大数据和互联网带来的跨界颠覆。腾讯率先抢获移动互联网的门票，促使电信运营商趋于管道化。在大数据时代，“前向免费圈数据，后向创新来变现”成为新的玩法，企业家和投资人需要重新审视产业和企业的投资价值、护城河意义与核心竞争力。

在以往传统的行业中，一般都只盯着行业对手，就像三星盯着苹果，联想盯着戴尔，百度盯着谷歌。然而如今是跨界竞争时代，跨界竞争常令对手防不胜防，企业尤其需要注意跨行业技术的颠覆，这常使竞争对手双方在遭遇跨界颠覆时都一片愕然。比如，数码相机对柯达胶卷的颠覆，导航系统的普遍应用对传统地图的颠覆，还有特斯拉开放专利技术也是对汽油车的颠覆。

随着大数据分析挖掘技术的发展，原本与股票行业毫无关系的社交网络及微博客服务的网站也开始横插一脚。如有学者利用OpinionFinder和“情绪状态量表”（POMS）这两种不同的情绪跟踪工具来分析Twitter上将近1000万条微博的文本。结果发现：在这个基于谷歌的POMS测量法中，“冷静、警惕、确信、重要、和善、快乐”这六种情绪里面，“冷静”是具有预测价值的。单靠“冷静”这一情绪指标就能预测未来3～4天道琼斯工业平均指数的每日收盘涨跌，准确率高达87.6%。利用大数据挖掘分析的舆情已经成为资本市场的“风向标”。

大数据时代，涉及最广、影响最深的行业里面，金融行业首推第一。易欢欢去年作了一份行业报告，该报告被誉为中国2013年影响力最大的一份行业深度报告：大数据时代的跨界与颠覆——金融业门口的野蛮人[②]。报告里详细分析了在大数据和互联网的冲击下，传统金融行业所遭遇的巨大挑战和面临的新对手、新模式。

伴随着大数据及互联网的迅猛发展与广泛普及，越来越多的互联网企业开始跨界涉足金融业，此期间涌现了大量的初创企业，给传统金融业的多个领域带来了较大的冲击。传统金融机构（银行、基金、保险、证券从支付结算到投融资服务，再到流通货币）等均无一能幸

---

① http://blog.sina.com.cn/s/blog_d0b1992e0101mu0w.html

② http://blog.sina.com.cn/s/blog_573968220101ases.html

免，渗入范围日益扩大，并开始向金融业的核心领域进行拓展。如，马云在2012年9月的网商大会上特别提出三大业务即平台、数据及金融，而后，在金融业务领域上，阿里巴巴展开了一系列的动作及布局。其实早在2011年4月，支付宝就与10家银行如工商银行、中国银行、农业银行和建设银行等联手大力推出“快捷支付”，有效打破了网银交易的额度限制，快捷支付摆脱了银行的束缚，不再需要对银行的网关进行链接，牢牢地抓了海量终端用户的消费数据，与此同时，还使得银行不得不与支付宝分享其核心客户的数据资源。消费者对于“快捷支付”的简便极其喜欢，这从仅推出7个月，用户数就已经突破两千万这一数据即可窥见一斑。截至2012年10月，与支付宝合作的银行机构已经超过了100家，用户数量更是超过了1亿。从一开始的网银通道，拓展到新型的业务如快捷支付等，支付宝对虚拟账户及用户数据进行截流，并把这紧紧抓在手中，支付宝已不仅仅承担着银行的渠道功能，开始实现被动与主动之间的转换，甚至慢慢成为主角。

同样，阿里巴巴的金融业务——阿里小贷，借助阿里巴巴（B2B）、支付宝、淘宝等多个平台长年积攒的大量数据，对小微企业的风险问题有效地进行控制，并通过互联网的流水化及批量化作业，大大降低了业务的成本。阿里小贷领先创造了从风险审核到放贷的全程线上模式，通过多维数据交叉验证、关联规则分析等技术进行风险评估，构成风险控制的双保险。同时，阿里小贷还凭借互联网技术监控贷款的流向，实现精细化管理，例如客户若将贷款用于扩展经营，阿里小贷将会对其广告投放、店铺装修和销售进行评估和监控。阿里小贷凭借着其独特的商业创新与竞争优势，在小微企业融资领域中快速地发展起来。根据相关数据显示，仅仅在2013年一季度，阿里小微金融放贷金额达到120亿元，有超过100万笔的发放贷款，平均单笔贷款额度大概为1.1万元。相似的是，美国的Lending Club公司对借贷人评分的主要根据是其社会安全号码、信用报告（报告内容包含三年的信用记录历史等）以及FICO评分（FICO分数区间为300~850，要求高于660分）等内容综合进行，为资金供求双方构建了一个便捷的大数据平台，金融交易与以往相比，其成本更低、而效率却更高，这不止扩大了服务人群的范围及数量，还加速了金融脱媒的进程，摒弃了对传统金融中介（银行等）的依赖。

又如，中国起步于2007年的P2P网络借贷，有着非常迅猛的发展态势。一批新兴P2P网络借贷公司如拍拍、红岭创投、宜信、贷帮等顺势而生，不少传统巨鳄开始在P2P市场上大施拳脚。但是，中国对于这个行业的监管政策目前几乎处于空白，高利贷、非法集资、庞氏骗局等违法行为有可能借助P2P网络借贷展开。从中国最严谨网络借贷平台——哈哈贷的轰然倒闭到不久前的“优易网”忽然跑路，“倒闭”事件与“卷钱跑路”事件在P2P网络借贷不时发生，揭示着该行业缺乏准入门槛和行业标准，故而市场上鱼龙混杂、良莠不齐，为该行业的发展蒙上了一层阴影。亟待解决的重大难题除了政策桎梏之外，另外一个就是信用风险。在国外能够轻松查找出个人完整的信息及其信用档案，而在我国，控制个人征信系统的是中国人民银行，能进行查询的唯有银行金融机构，并且这信息本身也不是很健全，这导致P2P网络借贷面临的风险将更大。实际上，大数据技术有望破解这一难题，其中有两项重要的任务：一是如何获得更多关于借贷双方的数据信息（特别是借款人）；二是如何很好地将数据信息融合到实际业务运作之中。美国Lending Club引入网络社交元素和中国宜信重点布局线下业务，均能对服务对象数据信息的来源渠道进行扩大。并且随着时间的推移，针对每一

个用户P2P网络借贷平台都会建立一个信用记录档案，相信伴随着国家相关数据的开放、数据来源的丰富以及平台数据的积累，我们将会逐步解决P2P网络借贷的数据获取问题。紧接着应该建立一套适用于大数据的业务模式，特别是在对应利率设置、信用风险评估和关键流程设计方面，仅仅利用大数据技术是不足以解决问题的，还需要企业立足于金融产业，深度地融合统一金融与技术[①]。

总而言之，大数据时代，中国金融业在新兴信息技术和国家大资管政策促使下，其竞争日趋激烈，而最终胜出的企业必然在金融与数据信息处理方面都具有强大的能力。

8.6.2 大数据时代，全新的投资理念和巨大的投资机会

随着大数据时代的到来，越来越多的企业将基于大数据资产进行商业模式的创新，甚至跨界涉足其他产业，并对该产业形成巨大的冲击。金融、电信、教育、医疗等各个产业未来都将感受到大数据的颠覆力量（图8.1所示是大数据颠覆的传统行业）。而在这个过程中将会衍生出两大新的投资的机遇：一是互联网企业传统产业化，大量互联网企业将基于大量的用户和数据资产进行跨界变现；二是传统产业互联网化，面对新进入者的冲击，传统产业中的企业将会积极进行自我变革，加大IT投入，进而带动相关IT服务企业的增长。现今，大数据的价值在金融领域尚无明显的体现，但大数据技术核心之一的数据挖掘是近年来极为重要的投资领域，随着大数据的发展，这项创新的分析广度及速度都会大幅度地提升。阿里金融已经在传统金融领域中，得到了行业的广泛关注。阿里金融低廉的信贷征信成本对中国现有的

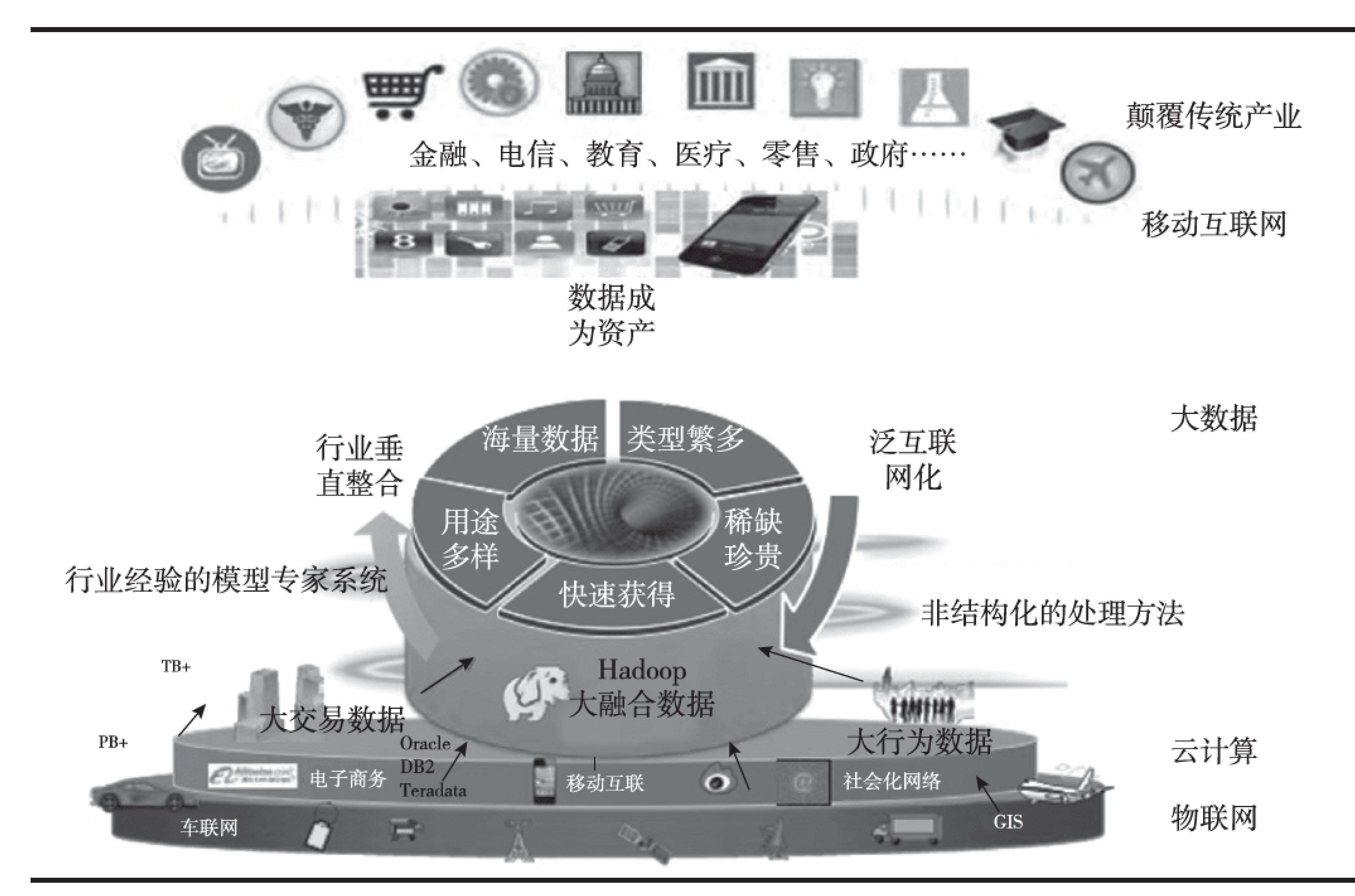


图8.1 大数据颠覆传统行业

① http://blog.sina.com.cn/s/blog_573968220101ases.html

金融机构和模式产生了巨大的影响。目前证券公司建立了的客户关系管理系统，该系统通过对客户交易行为进行分析，挖掘其风险偏好，从而进行合理的资产配置推荐，这个管理系统当前尚处于运用初期，相信在未来会有广阔的发展前景[①]。

大数据时代给我们带来了巨大的变化，网民和消费者的界限、企业的疆界都在逐渐变得模糊，数据俨然已成为核心的资产，并将对企业的业务模式造成深远的影响，甚至重构其文化和组织。可以毫不夸张地说，大数据将对国家的治理模式、企业的决策、组织及业务流程、对个人生活方式都将产生重大的影响。如果无法利用大数据进一步贴近消费者、深刻了解客户需求、有效分析信息并作出预测决策，传统的产品公司极可能沦为新型用户平台级公司的附庸，管理也无法逆转其衰落[②]。因此，新一轮信息化投资与建设热潮将被大数据引发。根据互联网数据中心预测，2020年，全球总共拥有的数据量将达到35ZB，麦肯锡则预测未来大数据产品在三大行业的应用中将产生七千亿美元的潜在市场，而中国大数据产品的潜在市场规模极有可能达到1.57万亿元，IT行业将迎来一个新的黄金时代。设备提供商与数据处理技术，ERP、BI、CRM改造服务商与IT系统咨询，信息安全提供商以及智能化和人机交互应用等都将获庞大需求，相应公司将获得巨大的机会。

鉴于国际巨头在硬件层和基础软件层具有明显的垄断优势，本土的企业将主要依赖于对客户需求的了解与客户资源优势以及本地化服务的优势，在应用软件层获利，而拥有大数据处理、挖掘技术、数据资产和数据分析人才的公司将在竞争中更具优势[③]。

未来的成功企业必然兼具传统产业和互联网技术与思想双重基因，谁能快速补足短板，谁就更有可能胜出。

8.7　练习

1. 统计预测的方法包含哪些？
2. 统计决策的概念是什么？
3. 大数据预测决策的关键点有哪些？
4. 在大数据时代下我国政府应如何迎接挑战？
5. 在大数据时代下政府公共管理水平的变革主要体现在哪些方面？

参考文献

[1] 拉里・罗森伯格，约翰・纳什，安・格雷厄姆. 大决策：大数据时代的预测分析和决策管理[M]. 上海：上海社会科学院出版社，2014.

[2] 徐国祥. 统计预测和决策[M]. 3版. 上海：上海财经大学出版社，2008.

① http://blog.sina.com.cn/s/blog_6030e2330102e60n.html
② http://news.xinhuanet.com/info/2013-03/20/c_132246718.htm
③ http://www.chinadaily.com.cn/hqcj/zgjj/2013-05-17/content_9062052.html

[3] 埃里克·西格尔. 大数据预测[M]. 北京：中信出版社，2014.
[4] 张武，袁其谦. 现代工业企业管理[M]. 北京：北京理工大学出版社，2011.
[5] 维克托·迈尔–舍恩伯格，肯尼思·库克耶. 大数据时代[M]. 盛杨燕，周涛（译）. 杭州：浙江人民出版社，2013.
[6] 徐继华，冯启娜，陈贞汝. 智慧政府：大数据治国时代的来临[M]. 北京：中信出版社，2014.
[7] 张才明. 数据驱动管理者决策[J]. 企业管理. 2013，（11）：110.
[8] 钱蓝. 时间序列法在市场预测中的应用[J]. 中小企业管理与科技（上旬刊）. 2011，（10）：33.
[9] 冯伟. 大数据时代信息安全面临的挑战与机遇[N]. 科技日报. 2013（001）.
[10] 陈美. 大数据在公共交通中的应用[J]. 图书与情报. 2012，（6）：23–24.

第9章

大数据与市场营销

随着海量数据时代的到来，数据的重要性日益凸显，对于市场营销而言，大数据为营销人员确定营销策略，量化营销效果提供了有力的技术支持，同时也在高管层为首席营销官赢得了一席之地。大数据技术将从各个层面改变传统的市场营销，迎来市场营销的全新时代。

9.1 大数据时代的营销模式创新

2011年，美国著名电视问答节目《Jeopardy》历史上最厉害的两位选手都被超级计算机Watson击败，而Watson也因此一举成名。Watson可以采用极快的速度处理数百万份以人类文字语言书写的文件。通常电脑只能轻松地处理传统数据库中的数据文件，但Watson的厉害之处在于它可以阅读网站中的信息、新闻报告、电子邮件等这类非结构性文件。为此，花旗银行花费巨额聘请Watson，借用它能够处理非结构性文件的能力来帮助花旗银行作出决定。比如应该对特定的客户提供哪些新的产品与服务，降低欺诈案件概率，以及搜寻哪些客户有信用度降低的迹象，这或许是Watson的第一份工作。

但这也仅是大数据时代的故事之一①。

9.1.1 营销模式的突出优势

数据生产及收集能力和速度的大幅提升掀起了“大数据”热潮，然而在营销领域，大数据的实践才刚刚开始。事实上，数字浪潮的到来，给消费者带来了直接而深远的影响。立身于数字的时代，消费者将更加独立，其自主意识逐渐增强，已经不再容易相信传统营销“轰炸式”的传播和灌输。相反地，一种新气象将会取而代之，消费者将会越来越喜欢质疑产品和品牌，喜欢在网络上发表个人的意见，这无形之中会给其他人造成一定的影响。此外，他们会愈来愈关注个人对品牌的价值，渴望品牌能够针对他们的意见和疑问迅速做出回应。若此时企业或厂商对于他们的观点采取漠视的态度，无疑会造成关注人群的大量流失，在这种时代环境下，传统形式的营销模式传播日显“疲态”。

以前，很多人都愿意相信这样一句话：中国拥有13多亿人口，即使一个人仅仅消费一元钱，那对于市场来说也是巨大的价值。可是如今，人们终于清晰而深刻地认识到，中国的13多亿人口拥有的消费模式会因为其收入、年龄、地域、文化以及生活方式等的差异而千变

① http://blog.sina.com.cn/s/blog_77217ab501017w76.html

万化。传统的营销模式给予营销人员两大坚定的信念——营销和品牌。可是，究竟是怎样的市场营销模式以及品牌传播策略才是真正有效的呢？相关调查显示，如今中国企业对营销和品牌的理解还不够透彻，这样带来的结果便是营销人员在实际工作中对这些理念的应用不到位，主要体现在企业营销理念的陈旧落后、企业专职营销部门设置的空缺以及营销人员专业知识技能的匮乏，营销管理技术和工具的缺乏等。无独有偶，近期《哈佛商业评论》中的一篇文章更加犀利直白地道出——《传统营销之死》，文章中提及涵盖广告宣传、品牌管理、公共关系以及企业传媒在内的传统营销手段已经失效。科技的日新月异，社交媒体环境的日益发展和壮大，传统的营销模式面对的无疑是一场声势浩大的革新，数字时代的滚滚浪潮将人们关注的焦点渐渐推向了“大数据”。

在大数据时代还没来临前，人们一般利用的是传统的营销数据，包括客户关系管理（Customer Relationship Management，CRM）系统中的客户信息、广告效果、展览等一些线下活动的效果。这些数据源提供的是消费者某一方面有限的信息，远不足以给出一个充分的提示和线索。现在，另外一批信息数据正逐渐占领了大众的视野，包括官方网站登录数据、地理位置数据、邮件数据、社交媒体数据等。这些非结构化数据，它们更多地以文字、视频或者图片的方式出现[①]，而且是层出不穷，来势汹汹。若将传统数据和非结构性数据进行对接，并且能够保持实时地更新，那么营销的游戏规则将会发生翻天覆地的变化，人们将迎来大数据时代营销模式的一场巨大变革。根据IDC（国际数据中心）和麦肯锡对大数据研究结果的总结，大数据主要能在以下四个方面挖掘出巨大的商业价值：对顾客群体细分，然后对每个群体量体裁衣般地采取专门的行动；运用大数据模拟实境，发掘新的需求和提高投入回报率；提高大数据成果在各相关部门的分享程度，提高整个管理链条和产业链条的投入回报率；进行商业模式、产品和服务的创新。在碎片化的网络世界，营销者需要在表象的分散和碎片背后，找到那些因兴趣或者共同的需求而重新聚集起来的东西，如果能捕捉到这种注意力，就会找到新的集中点。“大数据”就是这个趋势实现过程中的利器。

1. 更快、更低成本的数据采集

在社会科学领域里通常是采用抽样的方式来研究消费者。具体而言，就是遵循随机或者配额的原则去寻找消费者，并且采用调查的手段来获得数据。显然，这种方式对于小量的数据是适用的，然而，在大数据时代下，人们会依赖实时监测的方式或者追踪消费者在互联网上产生的海量行为数据，整个过程的优势就在于快速以及趋于零的成本。

2. 更完整的消费者描述

社会化的滚滚浪潮驱动传统互联网平台逐渐向社会化方向转型，和以前不同的是，消费者越来越喜欢在微博、论坛、社交网络上讨论产品和品牌，这些日积月累的数据对于营销者来说尤其珍贵。通过丰富的消费者数据，比如地理追踪数据、社交数据和网站浏览数据等，可以更好地完善消费者的信息。英国GSK公司就已经开始尝试定位那些谈论过旗下子品牌的人们，并且对这些人在公开论坛上所谈到的所有其他内容进行追踪，再据此建立消费者描述，并不断完善，然后合理地对营销部门已有的数据与获得的外部数据进行有效地整合，从

① http://blog.sina.com.cn/s/blog_77217ab501017w76.html

而设定更加精准的促销及优惠，吸引更多的消费者来到对应的子品牌网站[1]，这无疑会给该品牌带来更大的市场号召力。

3. 对原有营销方式价值的再次发掘

IBM的一个商业合作伙伴正在进行一项研究，该研究希望将呼叫中心产生的所有对话转换成文字，从而能够实现对这一营销渠道的数据挖掘。如果该项研究能顺利实行，则能够让市场部门轻易获得之前所没有的对消费者的洞察力，同时了解消费者对新产品的回应以及对品牌的感受。IBM新兴技术项目总监Peter Waggett认为更深入的数据挖掘能够帮助许多公司找到商业问题的解决方案。

4. 更精细的消费者细分

消费者细分并不是一个新鲜的概念。传统的营销多数是以人口统计学特性来归纳总结目标消费者的，但是诸如消费习惯、心理特征、兴趣爱好这样的数据则需要依赖第三方的市场调查公司。借助于大数据技术以及更好的分析工具，营销者可以无限接近，甚至近乎准确地判断每一个人的属性，从而对消费者进行细分，真正做到个性化划分，而不再是简单的划分群体。例如零售商Williams Sonoma将他们6000万的客户数据库与其家庭信息链接起来，根据这些家庭的收入、孩子数量及房屋价值等进行顾客精准划分，基于不同消费者群体的选择偏好和行为方式来对其电子直邮邮件进行设定。事实证明，根据这些信息的直邮邮件所获得的反馈数量是没有进行精准化之前数量的18倍[1]，这足以证明对消费者的精细划分所能产生的巨大价值。

9.1.2　营销模式的创新之举

事实上，以客户为中心的营销模式，在提法上早已不新鲜了，但问题的关键在于，企业所遵循的营销模式是否真的“以人为本”。事实上，我们已经非常明确这个新模式是什么样子，同时很多企业也已经开始使用这种模式。《哈佛商业评论》就曾总结出以下关键点[2]。

1. 恢复社区营销

实际上，只要能够恰当地运用社交媒体，购买者在当地社区寻求购买体验的趋势就会被加速。譬如说，当消费者想要购买一台全自动洗衣机，或者是想要打听一位医生时，选择去找销售员商量，抑或是到某公司网站上去了解内容显然是不可行的。一般情况下，很多人都会倾向于向邻居或是朋友打听——这属于个人的关系网络，通常可以在私人的关系网络里面获得自己满意的答案。

企业家们应当尽可能地将自己在社交媒体上的努力用于复制这些社区指向的购买体验。而与Facebook相似的社交网络公司则更应该把它作为自己的优势——通过扩大消费者自己的关系网络，让越来越多的人提供他们对某件产品或者服务的购买体验。例如，一家新成立的公司就能够让忠实的顾客们在社交平台上轻松愉快地帮助他们做免费宣传。该公司首先采取必要的调查，只要某个客户在接受调查时表示自己能够做“推荐人”，公司就会立刻给他们

① http://blog.sina.com.cn/s/blog_77217ab501017w76.html

② http://blog.sina.com.cn/s/blog_a6982e460101630r.html

发一张表格，邀请他们在某些社交网站上填写对产品的使用感受或推荐。当填写完成后，该公司再将它转移到某一指定网站，然后“推荐人”的交际圈很快就会知道他对该公司的看法，最终这家公司可以轻易地达到宣传的目的。

2. 找到可以影响客户的人

许多公司都将大量的资源耗费在追求外在的影响者方面，比如那些所谓的网络红人。其实相对于这种方式，去寻找和培养能够影响目标客户的人，再让这些能影响目标客户的人来帮忙做宣传将是更好的选择。实际上这里就涉及了一种全新的顾客价值观，它并不是仅仅基于客户的购买量，这是与顾客生命周期价值的不同之处。商家常会用金钱作为衡量客户价值的标准，殊不知除了金钱之外，这个标准的覆盖面应该是非常广的。例如，某个顾客受到多少的尊重，顾客的关系网络对公司的战略影响以及其影响力的大小等。微软的MVP（最有价值专家）成员中有一位名为“Excel先生”的成员，他网站的访问量有时候甚至会超过微软的Excel专栏服务界面中的访问量，据此不难推测出他对微软有着极其重要的价值。于是，微软开始进行一些尝试，例如给这位Excel先生透露一些内部消息，并且让他可以提前试用微软的一些新产品。而Excel先生和其他的MVP们则利用他们的影响力帮助微软开拓新的市场来作为回报，这在某种程度上给微软节省了很多成本。

3. 帮助他们建立社会资本

当然，新型社区导向型营销的实践者们可能会反思这种利用MVP的顾客价值主张。在以往，营销为了获得顾客的支持通常用的是金钱奖励、礼品、折扣或者其他小奖励，而这种新的营销方式却另辟蹊径，努力帮助拥护者和有影响力的人建立社会资本，比如让他们收获新的知识，提高他们的声誉，帮助他们建立社交关系等，这些都是“影响者”渴望得到的。某公司曾经与他们的顾客影响者——一个IT公司的中层经理合作，这是一种特殊的创造性方法，通过向他们提供财务证明与强大的研究，这些经理在向他们的高层展示这个公司战略优势的同时，也获得了能够接触最高管理层的机会。与此同时，IT公司的中层管理者的声誉也会随之提高，他们会被认为是可以为高层带来新启发的战略思想者，而这种双赢的战略效果正是双方共同期待的。

4. 让拥护你的客户参与你的解决方案

数年前，美国青少年吸烟的人数越来越多，达到了警戒值，为此佛罗里达州对他们为缓解该问题而做出的长达十年之久的努力重新进行了反思，人们认为没有什么会比说服青少年戒烟更困难的了，因此当时很多人对此事并不看好。然而佛罗里达州通过建立同伴之间可相互影响的社区，竟然解决了青少年吸烟这个问题。他们首先寻找一些学生运动员、领袖或者外表上酷劲十足的孩子，这些人在青少年中有很大的影响力。接着将这些人分为两类，一类是不抽烟的，而另一类是想戒烟的。之后，他们会请求这些学生参与并帮助想戒烟的学生，而不仅仅是传达一个信息[①]。这的确是一个引人注目的例子，虽然涉及的内容仅仅是青少年的吸烟问题，但是该方法的成功无疑证实了“让用户参与你的解决方案”的必要性和重要性。

---

① http://blog.sina.com.cn/s/blog_a6982e460101630r.html

9.2　大数据时代下的网络化精准营销

我们可以设想这样一个场景：当某个顾客进入店铺后，零售商搜索公司的数据库，如果发现这位顾客是有价值及其希望留住的顾客，那么零售商就可以通过顾客过去的购物历史与微博主页综合起来获得综合的信息，来了解需要花费多少钱才能留住这位顾客，进而确定所售卖物品的合适价格以及零售商能够退让的利润空间，从而针对这位顾客提供个性化的沟通方式和最佳的优惠策略。这使得营销者能够在一个恰当的时间和恰当的渠道将最合适的产品和营销提供给一个潜在用户或者老客户，这就是最终的精准化营销的实现①。

受经济全球化和全球信息化、人类社会发展和需求多样性、云计算和物联网技术深化应用等方面的影响，大数据已经成为IT领域和互联网上反复提及的热词。大数据时代的到来，使得广告主对精准营销的需求也日渐增长。市场经济的产物是营销，营销的目的就是给企业找到市场，通过营销活动给企业带来效益。伴随着3G时代的来临，信息产业的融合日益增强，传统的营销方式在一定程度上被精准营销取代，精准营销将逐步占据企业营销的主导地位，最终成为现代企业营销发展的新趋势。

9.2.1　精准营销概述

2005年，世界级营销大师菲利普·科特勒教授第一次提出了“精准营销”的概念。所谓精准营销，就是在精准定位的基础上，依托现代信息技术手段建立个性化的顾客沟通服务体系，实现企业可度量的低成本扩张之路。我国精准营销理论体系创建者徐海亮认为，精准营销是通过现代信息技术手段实现的个性化营销活动，通过市场定量分析的手段、个性化沟通实现企业对效益最大化的追求。按照普遍的观点，精准营销应具有三个层面的含义。第一，从意识上说，要有精准的营销思想。营销的终极追求是什么？是无营销的营销，而要达到这种终极的思想，其必须的过渡就是实现逐步精准。第二，在已经树立了精准营销的思想后，真正实现它，还需要借助实施精准体系的保证和手段，而这种手段是可衡量的。第三，达到低成本可持续发展的企业目标。与传统营销相比，精准营销具有更加突出的优势，见表9.1。

表9.1　传统营销与精准营销的比较

传统营销	精准营销
盲目地采用传单、广告等传统手段	借助先进的网络通信技术、数据库技术等科技手段
中间渠道杂多，成本较高	以现代高度分散物流为保障，降低营销成本
缺乏和客户沟通的手段	与顾客进行长期个性化沟通
局限于定性的市场定位	可量化、精准的市场定位
高成本	低成本

如表9.1所示，精准营销中，运用先进的网络通信技术与数据库技术等手段，能够在长期个性化沟通方面让企业和顾客之间达成共识，并且为保持企业和客户之间良好的沟通以及为企业建立稳定忠实的客户群奠定坚实的基础，最终能够满足企业长期、稳定以及高速发展的迫切需求。得益于现代高度分散物流的保障方式，企业能够摆脱杂多的中间渠道环节，并

① http://finance.qq.com/a/20120801/006484.htm

且脱离对传统的营销模块式组织机构的依赖，真正实现个性化关怀。可量化的市场定位技术（market test）的采用，使得精准营销打破了传统营销仅能用以定性的局限。此外，精准营销真正使企业营销达到了可调控、可度量的要求，同时改变了传统广告沟通所必需的高成本，使企业能够低成本地快速增长①。

9.2.2　网络精准营销模式

网络的发展为精准营销提供了更加广阔的平台。根据2012年9月在北美进行的一次互联网数据统计，在一分钟的时间里，全球的电子邮件用户共计发出了2亿封电子邮件；移动互联网会增加217名新用户；Google会处理200万次的搜索；消费者会在购物网站为电子数据支付27.2万美元。作为一种新的媒体，互联网给世界带来的转变不仅仅是字面上理解的精准，还驱动着事实上的精准。营销领域最活跃、最具创造力的部分将是基于互联网的“精准营销”。精准营销主要的重心是在网民的行为特征与消费心理的识别上，一般会从地方生活门户网站、专业性门户网站、专门的信息网站、Email、微信、微博以及搜索引擎网站上获取网民的特征及行为习惯，在此基础上对网民的消费意向进行推测，并充分挖掘其消费潜力，进而“投其所好”，针对性地对其进行特定的商业信息展现。大数据时代下的网络化精准营销正是在先进的数据库技术和网络通讯技术基础上发展起来的新型营销手段，具有十分光明而广阔的前景。那么，这种网络化的精准营销模式有哪些具体表现呢？曹利菊等在电子商务中实施精准营销中的策略观点如下。

1. 网络广告的精准传播

以往企业在进行电子商务时，采用的投放方式是传统的粗放式网络广告，然而这并没有给企业带来相应的效果。信息技术飞速发展的今天，随着网络的发展，网络广告形式也在日益丰富，不断创新，相继涌现了点告、窄告等新型网络宣传方式。这些宣传方式更加注重精确地选择目标客户，最终实现了网络广告的精准传播。

（1）点告。这是一种全新意义和全新形式的广告，从字面上理解，“点告”就是要以点而告之取代广而告之，改变传统的片面追求广告覆盖面的思路，转向专注于广告受众人群的细分以及受众效果。例如企业可以通过问答的形式把企业的产品向目标群体推广，而目标群体则根据回答问题的数量多少得到不同的利益，最终实现宣传的目的。一般受众人群如果能够注册成为点告网的用户，且在网上答题即可获得相应的积分，再用积分换取相应的金钱收入或礼品。具体的模式就是，在用户注册为点告网的用户时，填写自己的职业、爱好等资料，点告网就可以根据用户填写的个人信息进行推测，然后将相应的题目推荐给用户，继而对受众进行自动分组，进一步精确地区分目标用户。“点告”与媒体存在相似之处，以其精准性、趣味性、参与性及深入性，潜移默化地影响目标受众，最终达到宣传企业的目的。

（2）窄告。顾名思义就是与“广告”相对立，这是一种把商品信息有针对性地投放到企业想要传递到的那些人眼前的广告形式。基于又精又准这种精准营销的理念要求。当投放

① http://blog.sina.com.cn/s/blog_4b90e11d0100da79.html

广告时，采用语义分析技术对广告主的关键词及网文进行匹配，这样便可以有针对性地将广告投放到相关文章周围的联盟网站的窄告广告位上，即“窄”广告。此外，“窄告”还具有另外一大特色，即能通过地址精确区分目标区域，锁定哪些区域是广告商指定的目标客户所在地，仅在这些相应的区域中投放，最后成功地精确定位目标受众。由此可见，随着分众传播的思路，“窄告”通过与信息网络技术相结合，最终实现了网上的分众传播。

2. 精准的市场定位

现代营销活动中关键的一环便是市场的区分和定位，唯有准确地区分市场，方能确保有效的产品、品牌以及市场定位。市场定位要求从各种角度，如客户认知、客户需求以及竞争者等，来综合考虑企业提供的产品和服务所应当满足的客户群体。网络化的营销模式下，对客户或者消费者的行为进行精确的衡量和分析，是实现精准的市场定位的要求。此外，还要建立相应的数据体系，利用数据分析对客户进行优选，同时为了确定所做的定位是否准确有效，可采用市场测试验证来加以区分。客户关系管理（Customer Relationship Management，CRM）系统记录了客户或消费者的基本信息及其消费行为，对于企业而言这至关重要，可依据该系统中的数据，进行数据挖掘，制定营销策略。

3. 精准的营销服务

针对潜在的客户或者消费者，企业通过各种现代化信息传播工具，并提供一对一的沟通服务直接与客户或消费者进行沟通。而对于那些曾经有过交易记录的客户，因为已经拥有了客户或消费者的基本信息、已购买的产品甚至购物过程中所浏览产品的相应记录，利用这些数据资料，可以在一定程度上分析客户或消费者的消费习惯、需求及心理。此外，还可以追踪消费者在各种社交媒体平台上发表的意见，从而进行数据的收集和处理，通过分析推测客户或消费者现在需要的是什么以及可能想要的又是什么。另外，除了上述的直接沟通方式之外，企业在进行一对一的沟通服务时，还可以利用电子邮件的方式将分析得到的相关信息发送给客户或消费者，并追踪客户或者消费者的反应。

关于大数据的精准营销有这样一个故事。海尔的社会化客户关系管理（Social Customer Relationship Management，SCRM）会员大数据平台于2012年创立，这一大数据平台定位于把企业内部的全流程数据动态地打通，以用户最佳体验为导向，驱动产品数据、销售数据、供应链数据、服务数据等全流程数据的优化增值，同时与企业外部的全网络数据动态连接，最终形成全流程用户体验生态圈。2013年4月，海尔把探针伸入了SCRM会员大数据平台，提取出数以万计的海尔帝樽用户的数据，并与中国邮政的名址数据库进行匹配，建立了“look-alike”模型。这个模型可以将已经购买帝樽空调的几万名用户所在的小区分成几类，并打上标签。拥有帝樽用户的上海虹桥新城小区被打上了一系列标签，再把这些数据标签映射回中国邮政的名址数据库，找到有相似特点的所有小区，进而锁定了北京景泰西里小区。与此同时，海尔SCRM会员平台依据其自身优势同几家旅游、健康类杂志合作，为北京地区订户提供购买帝樽空调的优惠。该小区的陈某订阅了此类杂志，显然其比较关注自然、健康问题，而帝樽空调有去除PM 2.5系列的产品（PM 2.5指大气中直径小于或等于2.5微米的颗粒物，含有大量的有毒、有害物质，对人体健康和大气环境质量的影响较大），海尔由此预测，陈

某极有可能对帝樽空调去除PM 2.5的产品感兴趣。之后海尔给陈某投递了一封附上帝樽空调去除PM 2.5功能介绍的直邮单页，促使陈某直接去了实体店，经过一番体验后他便购买了一套。通过海尔的精准营销，陈某显然享受到了个性化服务。

2014年4月，家电零售企业国美电器与制造业龙头海尔联合开展了一次会员营销活动。在这场"年度会员超级购"活动中，双方为用户提供了涵盖短信、微信、电话等全触点的个性化营销方案，并通过大数据研究为不同用户推送不同的产品解决方案。海尔大胆的和成功的尝试无疑给后继者带来了强烈的激励和启发，各行各业也纷纷迎着这番强劲的浪潮大胆试水。这不是结束，只是开始，无数商家渴望以及期待着一场大数据时代下的网络精准营销爆发式革命的到来！

9.3 大数据应用与商业机会

"假如我们有了一个数据预报台，就像为企业装上了一个GPS和雷达，企业的出海将会更有把握。"马云在2012年网商大会上的演讲中形象地描述了数据的重要性。随着科技的发展，世界已经进入大数据时代，而这数据背后潜藏的巨大的商业机会更是让无数人跃跃欲试。以前只有Google、微软这样的公司能做大数据深挖，现在已经有越来越多的创业公司进入，不同的公司在不同维度的数据分析和服务方面正创造出新的商业模式。谷歌搜索、Facebook的帖子和微博消息使得人们的行为和情绪的细节化测量成为可能。挖掘用户的行为习惯和喜好，在凌乱纷繁的数据背后找到更符合用户兴趣和习惯的产品和服务，并对产品和服务进行针对性地调整和优化，这就是大数据的价值。如今大数据在越来越多的领域中逐渐得到广泛的应用，通过对大数据的存储、挖掘与分析，大数据在营销、企业管理、数据标准化与情报分析等领域大有作为。从实力雄厚的传统IT企业以及互联网公司到基于Hadoop平台的初创公司纷纷进入大数据领域掘金。

9.3.1 车载信息服务数据在汽车保险业中的价值

车载信息服务在汽车保险行业中的关注度非常高，该服务是通过汽车内置的传感器和黑盒来收集和掌握车辆的相关信息。可以配置不同的方案，使用黑盒来监测所有的汽车数据。还可以监测车速、行驶里程，以及汽车是否安装了紧急制动系统。车载信息服务数据可以帮助保险公司更好地理解客户的风险等级，并设置合理的保险费率。如果彻底忽略隐私问题，车载信息服务装置还可以跟踪到汽车去过的所有地点、何时到达、到达的速度以及汽车使用了哪些功能等。车载信息服务数据最初是作为一种工具出现的，它可以帮助车主和公司获取更好的、更加有效的车辆保险。能够期待的是，等到交通工具都安装了车载信息服务装置后，其他行业也可以开始使用车载信息服务数据，一旦车载信息服务开始大规模使用，人们以前所设想的某些场景将会得以实现。

可以大胆想象一下，未来全国有数以千万计的汽车都安装了车载信息服务装置，然后第三方研究公司可以通过匿名的方式为客户收集十分详细的车载通信数据。无论交通是否阻塞，也无论何时何地，这种数据反馈方式都会提供大量的车载通信信息。若研究人员能够掌

握大量汽车在每一个高峰时段、每一天、每个城市中的动向，他们就可以非常清楚地判断出车流产生的前因后果。当然，像诸如“如果某条路堵塞，堵塞会以多快的速度蔓延到其他道路”的问题也可以得到回答。车载通信信息的实现是交通道路工程师们梦寐以求的理想结局，它的出现将会给高速公路的管理模式带来翻天覆地的变化。在我们能够见证的未来将会是一番让人欣喜的光明前景。

9.3.2　RFID数据在零售制造业中的价值

无线射频识别（Radio Frequency Identification，RFID），亦称射频识别，是一种通信技术，可通过无线电讯号识别特定目标并读写相关数据，而无需在识别系统与特定目标之间建立机械或光学的接触。RFID读写器也分移动式的和固定式的，目前RFID技术的应用很广，如图书馆、门禁系统、食品安全溯源等。

RFID的最大应用之一就是制造业的托盘跟踪和零售业的物品跟踪。制造商发往零售商的每一个托盘上都有标签，这样就可以很方便地记录哪些货物在哪个配送中心或者商店。最终，商店中价格很低的商品也可以配备RFID芯片，或者使用某一种类似的新技术。此外，RFID的一种增值应用是识别零售商货架上是否有相应的商品。如果读卡器能够连续不断地确定货架上每种商品的数量，当需要重新配货的时候，商家就能得到准确的消息。此外，RFID标签还可以跟踪锁定储藏室中某商品的存货量。当确定货架上没有该商品时，商家们只需要完成把储藏室的商品搬到货架上这个简单的动作就可以了。当然，实现这个目标在成本和技术上还存在一些挑战，但是对未来我们应该持有乐观的心态和坚定的信念。另外，可以有这样一个有趣的设想，如果将RFID与其他数据结合起来，其发挥的作用将是无穷无尽的。如果公司可以将配送中心的温度数据收集起来，当出现掉电或者其他极端事件时，就可以跟踪到商品的损坏程度。仓库数据还可以和装运数据联系起来，当商品发生损坏时，公司可以有针对性地召回商品，并且通知零售商当收到商品时再次对商品实行开箱检查。与其他的大数据源类似，RFID数据本身并不能发挥所有的威力，但当和其他数据有效地结合使用时却能发挥巨大的作用。

事实上，RFID技术的应用十分广泛。除了已经讨论过的制造业和零售业，RFID标签的另外一种应用是贴在赌场用的筹码上，每一个筹码，特别是高价值的筹码都有自己的内置标签，依赖这样的标签，赌场可以通过上面的串行编号实现惟一地识别。RFID技术的应用表明，一些底层相同的技术可以支持不同的大数据流，这些大数据本质相同，但是范围和应用却大不相同。

9.3.3　大数据在医疗行业中的价值

华盛顿中心医院与微软研究中心合作分析了多年来的医疗记录，包括患者的统计资料、检查、诊断、治疗资料等。这项研究发现了在什么情况下一个出院病人会在一个月之内再次入院。除了疑难杂症外，如果最初的诊断中有类似“压抑”这种暗示心理疾病的词，病人再度入院的可能性就会大很多。虽然两者之间没有绝对的因果关系，但如果病人出院之后的干预是以解决病人的心理问题为重点，就可能会降低再次入院的概率，这样就可以提供更好的

健康服务。这些数据属于医院，微软只是提供了分析工具（Amalga系统）从这些数据中来发现有价值的信息。

医疗行业中产生的数据量主要来自于医学影像信息系统（Picture Archiving and Communication System，PACS）的各种医学影像、病理分析等业务所产生的非结构化数据。人体不同部位、不同专科影像的数据文件大小不一，因此PACS的网络存储和传输要采取不同策略。面对大数据的汹涌浪潮，医疗行业将面临前所未有的挑战和机遇。麦肯锡曾在报告中指出，除去体制障碍，大数据分析能够帮助美国的医疗服务业在一年创造三千亿美元的附加价值。表9.2以临床操作和研发为例，展示了医疗行业中大数据的潜在机会。

如表9.2所示，医疗服务提供方设置的操作与绩效数据集能够指导数据分析，同时创建可视化的流程图和仪表盘，促使信息透明化，从而极大地提高医疗的效率。远程病人监控装置则包含了血糖仪、家用心脏监测设备，甚至还有芯片药片——被患者摄入后，可以将数据实时传送到电子病历数据库。最后分析远程监控系统所产生的数据，从中能够提前预测病人的身体状况，降低急诊数量，有效地减少病人的住院时间，实现门诊医生预约量和家庭护理比例提高的目标。现在的临床决策支持系统可以达到分析医生的输入条目的目的，通过比较其与医学指引的不同之处，帮助医生避免一些潜在的错误，诸如药物过敏反应等。此外，大数据分析技术将使临床决策支持系统更智能、更出彩，这得益于其对非结构化数据分析能力的日益加强。比较效果研究（Comparative Effectiveness Research，CER）即通过全面分析大型数据集，如病人的体征数据、疗效和费用数据等，再对比多种干预措施的有效性，最终找出对特定病人具有针对性的最佳治疗途径。此外，医药企业在新药物的研发阶段，可采用数据建模与分析的方式，判断怎样的投入产出比才是最有效率的，进而配备最佳的资源组合，有助于加速医药公司向市场推进新药物的过程，生产出针对性更强的药物——有更高潜在市场回报和治疗成功率的药物。

表9.2　临床操作和研发中大数据的应用

临床操作	研发
医疗数据透明度	预测建模
远程病人监控	疾病模式的分析
临床决策支持系统	提高临床试验设计的统计工具和算法
比较效果研究	临床实验数据的分析

9.3.4　社交网络数据在电信业及其他行业中的价值

追踪一千个成员或用户并非难事，但是，这些成员之间的直接关联关系将会上升百万级别，而再考虑到“朋友的朋友”则关联关系会升至十亿级别，这就解释了为什么社交网络分析会是一个大数据的聚集所在。社交网络数据非常吸引人的一个地方是，它能够识别出客户能影响的整体收入，而不仅仅是他或者她个人提供的直接收入。

现在假设电信运营商有一个被评价为价值相对较低的用户，该用户只有基本的通话要求，并不会为运营商带来其他任何的增值收入。按照运营商以前的做法，只会根据该用户的个人账户情况来对其进行评价。在以前，如果这名客户通过打电话要求更换运营商，运营商

可能并不会考虑挽留这名在他们看来没有价值的客户。而在社交网络平台迅猛发展的今天，这名价值看似不高的客户，可能通过数据分析识别出，与该客户曾经通话的对象恰好是有着广泛交际圈和社会地位举足轻重的人物。有研究发现，一旦某位成员离开通话的圈子，其他与之有联系的成员也可能会随之离开，然后更多的成员也开始离开，就像一种连锁反应一样，最终的结果便是所有的成员雪崩般地离开，而这对于电信运营商来说显然是坏事一桩。因此，不难看出，对社交网络数据的深远应用是让电信运营商能够从其最初崇尚个体账户利益的最大化转向了客户社交网络利益的最大化，这对于整个电信业来说，也是一个不容错失的福音。当然，这里提到的这种方式除了适用于电信行业外，也同样适用于很多其他的行业。像执法部门也能够从社交网络分析中受益，在处理案件的过程中，识别出哪些人和问题人群或者问题个人存在联系，甚至有着间接联系。一旦发现某些人在多个地方出入，那么他或她就会被定位监控，这对于案件的侦破将会有极大的帮助。此外，社交网络数据分析能够让公司在识别有着庞大社交网络的客户之外，还能够锁定能够影响其品牌形象的地方。公司可以积极主动地鼓励客户参与到相应的社交网站，采取奖励的方式激励他们写评论和表达意见，一定可以取得不俗的效果。

9.3.5 遥测数据在视频游戏中的价值

遥测数据是通过传感器被遥测终端接收到的实时数据，它是视频游戏产业的一个术语，用来描述捕捉游戏活动的状况。在战争游戏中，遥测数据收集的是用哪种枪械开的火，在哪里开的火，向哪个方向开的火，以及枪械对各种东西的破坏程度。使用遥测数据，游戏制造商就可以了解到客户的私人信息，玩家实际的玩法以及玩家是如何与自己创建的游戏进行交互的精彩过程。与其他行业相似，客户的满意度在视频游戏中同样也是一个问题。视频游戏的独特之处在于要设置一条非常精彩的行进路线，游戏过程中要不断地设置关卡给玩家提供挑战的机会，但是挑战又不能太难，过于富有挑战性的关卡会让玩家满腹挫败感而放弃游戏，但是如果游戏关卡设置得过于简单，也会让玩家倍感枯燥而放弃。所以游戏分析显得格外重要，它可以给商家充分的提示，哪些游戏关卡可以让玩家轻松通过，而哪些关卡对于顶级玩家来说也很难过关，从而可以让游戏制造商根据分析的结果适当地调整关卡的难易程度，达到最终的平衡，这无疑会给游戏制造商带来更多的利益和价值。

此外，遥测数据还可以依据游戏风格将玩家分类。这些重要的信息可以给设计者带来更多的设计灵感，足以催生更多更加优秀的游戏，同时还可以交叉销售现有的产品。遥测数据能够了解到玩家的认知层次，基于此可以改变整个游戏行业。游戏产业如今已经开始使用遥测数据，相信在不久的将来这个领域将会得到长足的发展，而遥测数据分析的效果将会给游戏制作以及推广方式带来重要且深远的影响。

9.4 大数据时代的商业变革

在互联网行业迅猛发展的时代，互联网企业乃至传统企业的管理者均对海量的数据规模及其爆炸性的增长极为熟悉，然而，对不同来源数据交叉形成的“大数据”是否真的具有庞

大的潜在价值，却是抱着将信将疑的态度。实际上，大数据所拥有的附加价值将远超企业管理者们的想象，可以说大数据将在商业模式和决策管理上掀起一场变革。

数据化意味着人们要从一切太阳底下的事务中汲取信息，甚至原本以为是错误的信息，包括很多以前认为和“信息”根本搭不上边的事情。例如很少有人认为一个人的坐姿能表现出什么信息，但事实是，当一个人坐着的时候，他的身形、姿势和重量分布都可以量化和数据化。日本先进工业技术研究所通过在汽车座椅下部安装了360个压力传感器，来测量人对椅子施加压力的方式，将人体的臀部特征转化为数据，产生独属于每个乘坐者的精确数据资料。该系统可根据人体对座椅的压力差异来识别乘坐者身份，准确率高达98%。这些数据都将有可能孕育许多前景光明的新产业。人们可以利用这些拥有独属身份的信息数据研发汽车防盗系统，而通过汇聚这些数据，未来科技将有可能研制出行驶安全系统，以减少安全事故的发生，不得不说，对于驾驶员而言这将会是绝对的福音。大数据的兴起，让人们将更多的目光投向之前忽略的角落，同时也将更多的不可能变为可能，它正以不可抵挡之势悄悄掀起一场惊心动魄的商业变革，冲破传统思维的约束和禁锢，引领新型营销的崛起和腾飞，它带来的将是一场对产业链有着强烈冲击的伟大变革。

9.4.1 大数据时代商业思维的变革

“大数据”的关键在于发现和理解信息内容以及信息与信息之间的关系，然而直到现在，人们对此还是难以把握。IBM资深的大数据专家杰夫·乔纳斯提出要让数据“发声”。按照人们传统的思维习惯认为，与数据交流存在困难是极为自然的，但却并未想过实际上这只是受其技术条件影响的一种人为限制。

在过去，由于受技术和资源的限制，社会学家、经济学家以及商人们，只能通过采样调研和统计分析手段去了解关注的对象，即在小数据时代，人们采用的是随机采样的方式，实现以最少的数据获取最多的信息的目的。然而，随机采样毕竟是在受限的时代应运而生的，现今的数据处理技术已然发生了翻天覆地的改变。当数据收集和储存的成本几乎可以忽略不计时，人们完全可以通过收集全面而完整的数据来加以分析。数据量的大幅增加会造成结果的不准确，一些错误的数据也会混进数据库，然而，在不断涌现的新情况里，允许不精确数据的出现已经成为一个新的亮点，而非缺点。因为放松了容错的标准，人们掌握的数据也多了，而且还可以利用这些数据做更多的事情。此外，人们还需要与各种各样的混乱作斗争。假设这样一个场景，要测量一个果园的温度，但是整个果园只有一个温度测量仪，那就得保证这个测量仪是精确的而且能够一直工作。反之，如果每100棵果树就有一个测量仪，那么有些测试的数据可能会出现错误，也可能会更加混乱，但是众多的数据整合在一起就可以提供一个更加准确的结果。这就是大数据应用容错性的突出体现——全部数据的采用会将有瑕疵的若干数据淡化处理。在大数据时代，新的分析工具和思路为人们提供了一系列新的视野和有用的预测，让人们将关注的焦点聚集到以前不曾注意到的联系上。一个重要的转变就是，人们学会了去探索“是什么”而不是“为什么”，不再局限于去验证已有的推测是否正确，不再是去仅仅寻找现象背后的因果关系。当然，在传统的思维里，快速思维模式使得人们习惯用因果关系来看待周围的一切。例如父母经常会这样教育孩子，天冷时不戴帽子和手套就

会感冒，然而，感冒和穿戴之间究竟存在直接的联系吗？答案显然是否定的。这正是小数据时代人们的思维模式，但是如今，大数据之间的相关关系将会用来证明人们在第一时间内直觉的因果联系是错误的。当然，因果关系仍然还是有用的，但是它将不被看成是意义来源的基础。在大数据时代，虽然在很多情况下，人们依然指望用因果关系来说明所发现的相互关系，但是，因果关系也只是一种特殊的相互关系。相反，大数据推动了相关关系分析，相关关系分析通常情况下能取代因果关系而起作用，即使在不可取代的情况下，也能指导因果关系起作用。维克多将大数据时代人类的思维革命总结成三个：不是随机样本，而是所有数据；不是精确性，而是混杂性；不是因果关系，而是相关关系。

9.4.2　大数据时代管理的变革

大数据所具有的庞大价值将影响到行业的各个枝节，毫无疑问会在商业模式及管理决策上掀起一场深刻的变革，而企业的管理层要首先转变思维，对企业管理模式进行变革。大数据时代下的企业要想在竞争中屹立不倒，既要掌握更多更好的数据，更要领导者有足够的管理能力，同时拥有先进的管理体系。那么大数据时代企业管理的变革又将体现在哪些方面呢？

1. 创新和发展的策略

有人可能会说，大数据时代下，管理层人员的视野、经验以及直觉在决策方面所起的作用会越来越小。然而事实并非如此，这个时代所需要的正是那些能够发现商机、开拓市场的领导者，能够具有创新锐利的思维且能让员工全心投入到该新想法的商业领袖，同时能够针对企业众多管理决策做出变革。这样的领导者将是未来十年能够推动企业成功的重要保证。

2. 广泛的实时用户定制

大数据提供了实时个性化定制的可能性，同时在用户定制上实现了质的飞越。在大数据时代，一切传统商业模式都会被个性化所颠覆，未来商业模式发展的新驱动力和终极方向将会是个性化。给个性化商业应用提供可持续发展的沃土与足够的养分是大数据的重要作用。可以想象，未来的商业可以实现实时用户定制的目标，即通过研究分析描述消费者个体行为和爱好的数据，为其提供专属的个性化产品和服务。这样完美切合的服务也是消费者所迫切向往的。

3. 数据跨职能部门的流动

大数据时代，企业面临的一个突出挑战，即高效的企业需要根据不同的部门把信息和决策分配下去。但是创造出来的数据信息究竟该运用在哪个部门？这便要求企业的组织架构应该极其灵活，可以最大化地进行企业跨职能的合作。除此之外，企业领导者需要为各部门的决策人员提供懂得相关技术的数据科学家，且给他们提供给合适的数据。

4. 充分理解大数据处理技术

近年来，对大规模、多形式数据处理的技术及工具有了非常大的改变。一般而言，这些技术与工具都不算十分昂贵，事实上，其中不少都是免费开放的。其中最常用的就是Hadoop，一个分布式系统基础架构，该架构能在大型、廉价的硬件设备上运行应用程序，同样在其平台上也提供了用于数据分析的工具。当然，这些新的技术和工具将给很大一部分企业

的研究部门带来新的挑战，比如对企业内部和外部数据的整合能力提出了更高的要求。总的来说，数据技术在大数据战略中并非占据最重要的地位，但是对它的需求却是必不可少的。

5. 对数据技术人员的管理要求

大数据的时代的到来使得企业对数据技术人员的需求日益增强，数据技术人员的价值也日益显著，而能够处理大数据的“数据科学家”更是极为重要的。就数据科学家而言，必不可少的技术之一便是统计技术，而传统的统计课程内容有局限性，处理大数据的核心技巧并不能从中学到。相对而言，清理数据以及组织大型数据的能力是大数据时代的统计技术更应关注的，因为海量的大数据里面除了有传统的结构数据，更多的是非结构化数据。最理想的数据科学家应是那些既能了解“商业语言”，又熟知大数据技术，可以从数据的角度帮助领导者做决策的专业人士，而这样的人才是各企业都千金难求的。

大数据的商业化应用使得人们的隐私问题和相关安全问题也日益凸显出来，现有的法律、法规已经无法保障这类安全的需求，管理者急需制定符合大数据模式的制度。在笔者看来，大数据的价值很美，但是不能让这些数据取代人们的自由选择和决策过程。比如通过大数据挖掘可以发现某人有潜在的犯罪倾向，但不能因此就认定这个人会犯罪，因为还需要考虑到道德主体的自由选择和环境的影响，或许选择适度的干预才是合理的。同时，那些尝到大数据甜头的人，可能会把大数据运用到它不适用的领域，甚至对大数据分析结果的依赖会过分膨胀。随着大数据预测的崛起，有些人对于大数据的崇拜会越来越明显，最终坠入盲目的漩涡。事实上，人们应该时刻保持理性的思维，警醒自我，始终坚信大数据并不是无所不能的。

9.4.3 大数据时代营销的变革

众所周知，当今世界经济急速增长，殊不知计算机数据处理能力的增长速度却是其增长速度的9倍，社交网络的全球扩张正是得益于这种急速的数据增长。毫不夸张地说，营销规则的命运将会因为这种增长而改写。在这个信息爆炸的时代，人们从在线社交生活中所发生的行为轨迹及其创造的内容使企业可以轻易地洞察客户的需求，这无疑是新型且高效营销的一个前提。亚马逊就基于从顾客身上获得的大量数据来分析他/她的购买习惯和偏好，并根据大数据的分析来构建了个性化的推荐系统及畅销书排行榜，这种对消费者的洞察力是大数据本身带来的技术解决方案，而不再是源于人工分析。

企业往往可通过公开的来源、网上数据库，通过购买与专门渠道来获得消费者数据，还可从网络社区和智能设施中收集各种新的资料。如今能够轻而易举地得到收集、分析数据的技术，且价格也在下降，这就造就了一个光明的前景——企业可以更进一步地深化对数据的使用。除此之外，数据应用已被很多企业逐步提升到一个新的层次，对推动在市场营销领域、实现彻底的变革有着至关重要的作用。在企业的营销活动当中，数据的采集、存储、分析挖掘与应用对营销决策有着至关重要的作用，所谓数据库营销就是对客户资料进行数据采集、存储、处理后建立的一个客户数据资料库，据此准确地对市场进行细分与定位，从而实现个性化、创新性的营销策略。因而，企业数据库营销的思想与如何在大数据时代有效利用大数据极其吻合，这也使得数据库营销成为大数据时代引领企业市场营销变革的主流方式之一。

每个企业赖以生存和发展的基础是“深刻洞察和理解客户需求”，而唯有对大量的客

户进行数据挖掘和行为分析才可能达到“理解”和“洞察”的目的。营销人员清楚地知道，市场营销实际上就是管理决策的过程，首先采用市场调研、客户群细分及定位等方式建立营销战略组合，而企业的战术组合则由产品、价格、促销以及分销相结合而成，同时运用有效的措施保证营销计划的实施。考虑到要提高营销的效果以及降低营销成本，企业应当具备针对性的精准营销，而数据库营销以一种适应大数据潮流逐渐成长起来的模式渐渐成为最佳的选择。数据库营销制定出“最易打动顾客及潜在顾客；与顾客建立起长期、高品质的良好关系；做到在适当时机以适当方式将必要的信息传达给适当的顾客、有效地抓住顾客的心思，让营销支持更有效益、建立忠诚度、增加利润”的营销模式，这样一来，可以为建立精准营销和良好客户关系奠定坚实的基础。随着信息技术的迅速发展，数据库营销成为适应现代信息社会和大数据时代的独特营销方式。比如在ZARA店内，当客人向店员反映“这个衣领图案很漂亮”“我不喜欢口袋的拉链”这些微小的细节时，店员即向分店经理汇报，经理通过ZARA内部的全球资讯网络，每天至少两次传递资讯给总部设计人员，总部根据汇集的数据做出即时决策后立刻传送到生产线，改变产品样式。ZARA通过收集海量的顾客意见而做出生产销售的决策已大大降低了存货率，同时根据这些数据分析出相似的“区域流行”，在颜色、版型的生产中，做出最靠近客户需求的市场区隔。

9.4.4　大数据时代产业链的变革

产业链是产业经济学中的一个概念[①]，是各个产业部门之间基于一定的技术经济进行关联，它包含四个维度的概念，分别是：企业链、价值链、空间链和供需链。这四个维度在相互对接的均衡过程中形成了产业链。产业链形成的内模式便是这种“对接机制”，作为一种客观规律，它犹如一只“无形之手”调控着产业链的形成。产业链的本质是用于描述一个具有某种内在联系的企业群结构，它是一个相对宏观的概念，存在两维属性：价值属性和结构属性。产业链中大量存在着上下游关系和相互价值的交换，上游环节向下游环节输送产品或服务，下游环节向上游环节反馈信息。下面以服装产业链为例来说明，见图9.1所示。

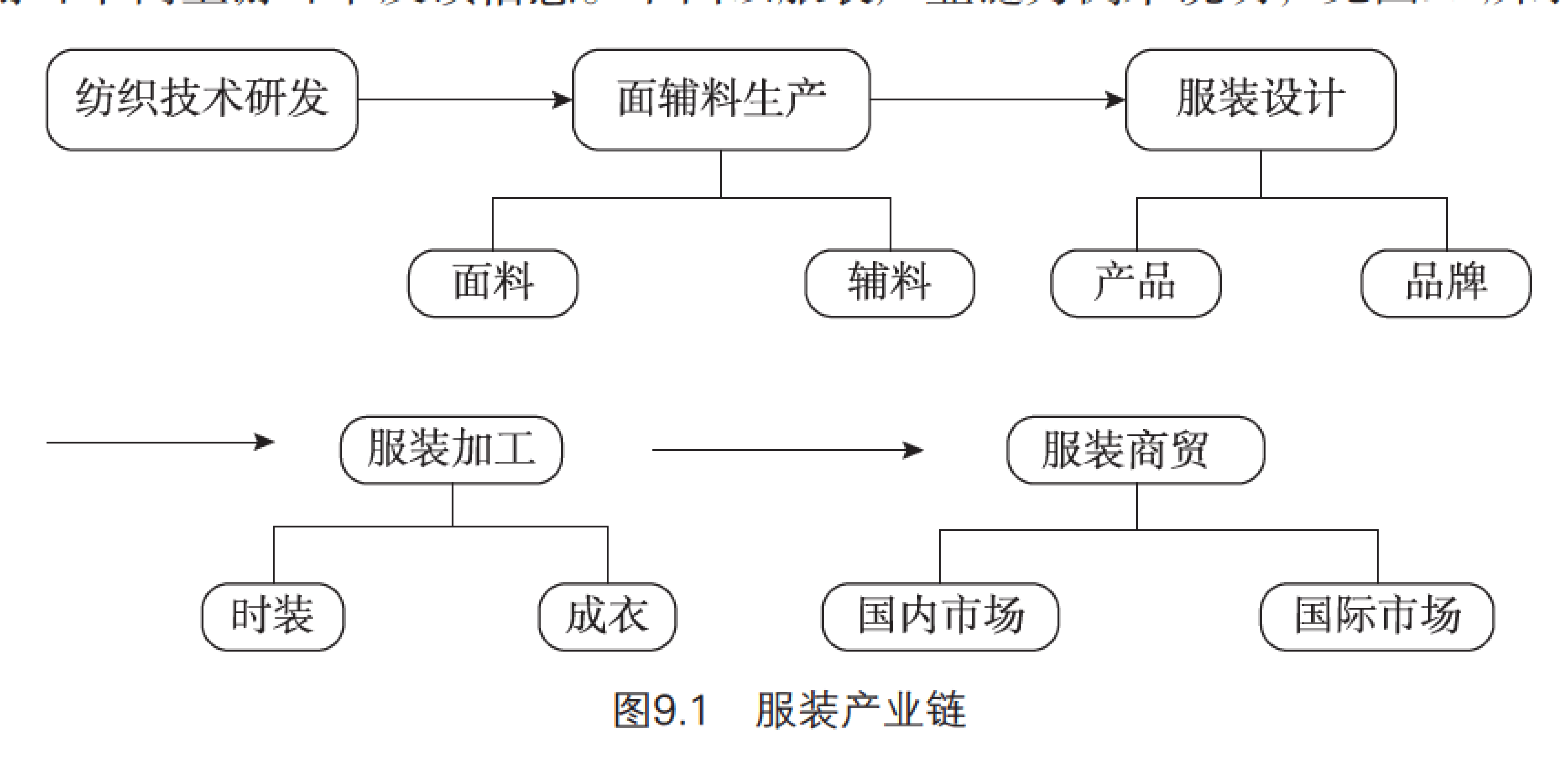


图9.1　服装产业链

权威机构预测，2014年全球大数据技术服务支出将超过140亿美元，2018年市场年均复合增长率将达26%，支出增至460亿美元。在第十二届全国人民代表大会第二次会议上，国务院总理李克强做政府工作报告时指出，要加快产业结构调整，鼓励发展服务业，支持战略性新

① http://baike.baidu.com/view/479661.htm?fr=aladdin

兴产业的发展，设立新兴产业创业创新平台，在新一代移动通信、大数据等方面赶超先进，引领未来产业发展。由此可以预见，数据服务将加快走向普及，并且需要一条成熟的大数据产业链给予支撑。麻省理工学院有利用手机定位数据和交通数据建立城市规划的先例；Tesco PLC(特易购)也开始利用大数据提高其运营效率，这家连锁超市在其数据仓库中收集了700万台冰箱的数据，通过对这些数据的分析，它可以进行更全面地监控并达到主动维修以降低整体能耗的目的。大数据产业链正是从囊括生产、科研、工业等广泛领域所产生的经典案例中不断成长并逐渐走向成熟的，它所发挥的重要作用就是能够从各种类型的数据中快速获得有价值的信息。在当前新的经济模式下，一些先行企业和机构的探索实践，使更多的人看到“数据”作为一种资源在产业化过程中如何发挥着它至关重要的作用——从生成、采集、传输、质量化、挖掘，最后形成“数据服务”对其他行业进行价值输出。目前，大数据产业链上活跃着三种类型的大数据公司。

- 基于数据本身的公司（数据拥有者）：拥有大量数据或者至少可以收集到大量数据，却不一定有从数据中提取价值或者用数据催生创新思想的技能，例如社交网络和运营商。
- 基于技术的公司（技术提供者）：技术供应商或者数据分析公司等。
- 基于思维的公司（服务提供者）：挖掘数据价值的大数据应用公司。

大数据的兴起对于激活产业链有着划时代的意义。阿里巴巴的金融业务就是基于其数据资产进行商业创新的重要成果。阿里巴巴在金融业的全面布局已经对传统银行、保险、小贷等多个行业形成了冲击，而在技术、模式和思维上已经形成较大的冲击，且将推动金融产业格局的重构，这也正是马云宣称要“摇一摇”传统金融的基础。大数据产业链及其技术体系的日趋成熟和完善是有目共睹的，相信大数据时代缔造的美好明天即将到来。

9.5 大数据提高企业竞争力

什么是企业竞争力？企业竞争力是在竞争性市场条件下，企业通过培育自身资源和能力，获取外部可寻资源，并加以综合利用，在为顾客创造价值的基础上，实现自身价值的综合性能力。同时也是在竞争性的市场中，一个企业所具有的，能够比其他企业更有效地向市场提供产品和服务，并获得赢利和自身发展的综合素质。企业的竞争力分为三个层面，第一层面是产品层，包括企业对产品生产及质量的控制能力、企业的服务、成本控制、营销、研发能力；第二层面是制度层，包括各经营管理要素组成的结构平台、企业内外部环境、资源关系、企业运行机制、企业规模、品牌、企业产权制度；第三层面是核心层，包括以企业理念、企业价值观为核心的企业文化、内外一致的企业形象、企业创新能力、差异化个性化的企业特色、稳健的财务、拥有卓越的远见和长远的全球化发展目标。总而言之，第一层面是表层的竞争力，第二层面是支持平台的竞争力，第三层面是最核心的竞争力。企业竞争力的强弱无疑是企业经营者们关注的焦点，那么在数据爆炸的时代，企业竞争力究竟发生着怎样颠覆性的变化？这场大数据革命会让企业经营者们面临怎样的挑战呢？

在以往漫长的商业社会进化过程之中，人才对企业有着至关重要的作用，离开了人才，企业的智商就基本为零，也正是由于这样，人才显得极其重要，甚至一度以企业的核心竞争

力而存在。一方面，这些人才的大脑中分布存储了企业的智商；另一方面，人才的商业智商能够借助于企业，帮助提高企业智商。在某种程度上，人才的商业智商完全决定了企业智商的高低[①]。而由于企业智商是分布存储在各个人才的大脑之中的，要进行信息的分享或是价值挖掘都会受到非常大的限制，在一定程度上往往很难完全发挥其作用，这显然不利于企业的发展。在大数据时代，传统观念里备受推崇的人才固然重要，却并不是企业智商最为重要的载体，如今企业智商真正的核心载体是海量的数据。那些能够按企业需要随时获得的数据，能够对企业业务流程的任意环节进行帮助和指导，利于有效运营和优化，并且能够协助企业做出正确合理的决策，这对于推动企业优化管理无疑是雪中送炭。大数据时代海量繁杂的各种数据的存在使得企业能够真正掌握企业全部智商。大数据好似洪水决堤般倾泻而来，如果把浩瀚的数据比作无际的海洋，那么企业必须如鱼儿熟悉水性一样，自如地熟悉、运用好大数据，才能真正做到游刃有余。除了企业智商被大数据重新定义以外，企业核心资产也同样被大数据重塑。在过去很长的一段时间里，土地、流动资金及人才等几个要素是用以衡量企业最重要的资产，而现在，数据将作为企业的一项更为重要的资产与企业的发展潜力有着直接且密不可分的关联。可以预见，大数据的洪流给企业经营者们带来的会是多么强烈的冲击。当重塑企业智商以及核心资产完成以后，数据资产毫无悬念将地成为现代商业社会的核心竞争力。

基于其行业的特殊性，大数据给商业所带来的巨大变化早已被互联网行业所感受到。不少互联网企业早在许多企业还在对大数据引起商业世界的变革无所适从时，就已经重新定义了其核心竞争力。在某种程度上，企业在大数据时代即将迎接的美好未来正体现在这些互联网企业目前所经历的巨大变化。亚马逊长期以来都采用大数据分析，试着对客户进行定位，获取客户反馈。亚马逊的CTO Werner Vogels曾经说过“一旦进入大数据的世界，企业的手中将握有无限可能”。然而，亚马逊在大数据时代做出的大胆跨越并非仅有，如今电商之争可谓诸侯争霸，火药味甚浓。大数据所爆发出来的巨大潜力，正如利剑出鞘、铠甲上身，必将在未来为电商企业的精准营销带来融合性的影响。古语有云：长袖善舞，多钱善贾，意指有所依靠，事情容易成功。可以预测，将数据转化为竞争力，将会是电商企业在大数据时代提高其核心竞争力的必经之路。通过对大数据的深度分析，企业能够把握市场未来的发展方向、消费者的采购行为以及企业营销的增长，这在企业竞争中成为不容忽略的巨大优势。过去，占领市场靠的是企业家的市场敏感度，今后则更多依赖于对数据的可控分析，特别是来自于互联网平台的海量数据。面对成千上万的大数据，电商企业要借助数据的力量，特别是互联网平台数据去支持电商企业的发展，需要创造性地运用互联网思维，把经过分析、挖掘后的互联网数据运用在电商营销中，所谓找到“买奶粉的刚需客户”，成功地把商品推送到需求客户身边。可以预见，在未来大数据分析与电商行业的结合会出现以下的趋势。

第一，在互联网平台端，来自用户的消费习惯、兴趣爱好、关系网络以及整个互联网的趋势和潮流都将成为电商行业从业者关注的热点，而这一切的获取和分析都离不开对大数据分析。

---

① http://content.businessvalue.com.cn/post/6687.html

第二，基于移动互联网与移动社交平台的海量数据分析，将电商营销带入了个性化时代。应用互联网平台的大数据分析，可以提示电商企业正确的营销时间与方向，潜在的用户和交易机会等，这正好切中了企业在移动互联网端的需求。

第三，根据对来自互联网等各类不同平台的数据进一步的挖掘和分析，找到这些数据相对应的人群，再将这些群体进行个性化的对比，并以此展开电商企业个性化的产品营销服务。

除了电商企业，蓬勃发展的大数据也向各行各业伸出了橄榄枝。不论是传统的石油行业还是传统银行业抑或是零售业，都意识到数据的重要性。据媒体报道，埃克森美孚曾在此前一次全球性招标中，一次性投入10亿美元来采购信息化服务。传统的商业银行也努力和互联网“合作共赢”，并进行模式创新，如推出POS网络商户贷款业务。全球最大的零售商沃尔玛也在其社交基因组计划中整合了用户在社交网络中的关系数据，用以更精准地推测消费者的偏好，以此来实现更加精准的营销。除此之外，大数据也向小数据时代的赢家以及那些线下的大公司，包括宝洁、联邦快递、雀巢以及波音公司等，提出了挑战。这些赢家也必须意识到大数据如今的威力，然后有策略、针对性地收集和使用数据。诸如科技创业公司和新兴行业中的老牌企业也并不例外，苹果公司进军移动手机行业就是一个非常典型的例子。在iPhone推出之前，苹果公司在与移动运营商签订的合约中明文规定，运营商们要提供给它大部分的有用数据，正是因为拥有多个运营商提供的大量数据，苹果公司所掌握的关于用户体验的数据比任何一个运营商都要多。有人可能会有这样的疑问，是否只有这些大型集团公司才有能力进行数据挖掘呢？中小型企业由于资金相对受限，是否就意味着在这场大数据革命中真的没有一席之地，甚至完全失去竞争力呢？在中小型企业如雨后春笋般发芽繁荣的今天，面对资金的限制和市场的竞争，小企业该如何更好地去迎接大数据时代的挑战呢？

事实上，大数据给小公司带来的不是危机，而是机遇。因为聪明而灵活的小公司能享受到非固有资产规模带来的好处，这也就意味着，它们并没有很多的固有资产却具有强烈的存在感，并且它们也可以通过降低成本的方式传播创新成果，这样也能取得不错的效果。还有一个关键所在，因为最好的大数据服务都是以创新思维为基础的，所以它们不一定需要投入大量的原始资本，而是应该把握好创新的原则。数据可以授权但是不能被占有，数据分析能在云处理平台上快速而且低成本地进行，而授权费用则应该从数据带来的利益中抽取小部分。换言之，传统的资源比如土地，几乎已经被瓜分殆尽。而数据资产却还处在跑马圈地的阶段，任何一个小企业都可以拥有这样的机会——通过提供新颖的服务来获取不同的数据资产，而大数据恰恰又可以给小企业提供后来居上的机会。例如，一家专门提供包车和租车服务的商旅运输公司，正常情况下是竞争不过传统出租车公司的，但是如果可以获取在线叫车服务乘客以及司机的双向数据，然后再根据这些数据针对不同客户的需求提供个性化的服务，便可就此实现超越的目的。因此，对于小企业而言，需要领悟到大数据的精髓和深层价值才能提高竞争力，进而赢得持续的发展和进步。

大大小小的公司都能从大数据中获利，这个情况很有可能并不只适用于使用数据的公司，同样也适用于掌握数据的公司。大数据拥有者会绞尽脑汁地想办法增加它们的数据储存量，因为这样能够达到以极小的成本换取更大利润的目的。首先，它们已经具备了存储和处理数据的基础；再者，数据库的融合能够带来特有的价值；最后，数据使用者如果只需要从

某个人手中购得数据，那将更加省时省力。不过实际情况要远远复杂得多，因为可能还会有一群处在另一方的数据拥有者的诞生，这对于已经存在的拥有者来说必然构成威胁。随着数据价值的显现，很多人会想到以数据拥有者的身份大展身手，他们收集的数据往往是和自身相关的，比如他们的购物习惯、观影习惯、外出吃饭的习惯，也许还会有医疗数据等。消费者可以非常自主地做出决定，如可以选择将这些数据中的部分授权给哪些公司。当然，除此之外，很多人也是愿意免费提供这些数据来换取更好的服务，例如那些想得到图书网站更准确的书籍推荐的人。现在，无论是消费者授权还是公司从个人手中购得信息都还过于昂贵和复杂，所以可能会催生一些中间商，他们会从消费者手中购得信息，然后卖给公司。如果这个过程的成本足够低，并且消费者对中间商也有足够的信任，那么个人数据市场的诞生就是可能的。

毫不夸张地说，大数据给人类带来的巨大商业价值丝毫不逊色于二十世纪计算机革命造成的巨大变革。那么在这场大数据洪流中企业能否抓住这个千载难逢的机遇，为提升企业的核心竞争力做出大胆的尝试和努力呢？为此我们给出以下几点建议。

首先，提供数据交易、迁移、存储、处理以及分析的实时平台，满足行业用户在大数据的突出挑战下快速、实时地处理和服务需求。

其次，将大量结构化与非结构化的数据进行整合处理，并且融合云计算应用程序，将其集成到电脑及各种工程系统中，从而使用户的工作简化。

再次，打造大数据优化解决方案，在确保数据真实性的前提下，有效处理大规模、多样化、高速流动的数据，帮助用户获取对其业务的理解和洞察力，用以制定相应的策略以及实现业务的快速突破和增长。

最后，创建数据的管道化管理流程，以数据集聚为依托，以各种数据应用为驱动，以丰富的界面形式对用户展现数据分析结果，最终完成数据的汇总、应用分析及结果呈现的完整流程。

如今的大数据时代，在不经意间给商业的生态环境带来了翻天覆地的变化[①]。网民与消费者的界限逐渐变得模糊，网络传输的随时在线，智能终端的无处不在以及社交网络的频繁互动都让往日仅仅浏览网页的网民逐渐变得“轮廓分明”。大数据使得企业第一次有机会对消费者行为进行大规模的、精准化的研究。而其中对变革具有向往意识的企业，更应主动地拥抱这种革新、变化，最终实现从战略到战术层面的自我的蜕变与进化。

9.6　练习

1. 大数据时代营销模式的优势体现在哪些方面？

2. 什么是精准营销？精准营销与传统营销的区别是什么？大数据时代下的网络化精准营销的具体表现有哪些？

3. 举例说明大数据时代下涌现的商业机会。

4. 简要阐述一下大数据时代的商业变革。

① http://content.businessvalue.com.cn/post/6687.html

参考文献

[1] 李海岚，蔡国良，冯宗智. 简营销：大数据时代市场营销的逆向思维[M]. 北京：机械工业出版社，2014.

[2] 冯利芳，麻震敏，孙珺等. Big Data大数据重塑营销[J]. 成功营销. 2012，40–41.

[3] 廖波，唐钢，陈娉婷. 大数据时代市场营销模式变革思考[J]. 中国经贸. 2013，66–67.

[4] 曹利菊，刘俊斌. 精准营销在企业电子商务中的应用研究[J]. 商场现代化. 2009，（05）：149–150.

[5] 陈秋阳. 基于数据挖掘技术的精准营销系统的设计与实现[D]. 杭州：浙江大学计算机学院，2010.

[6] Bill Franks. 驾驭大数据[M]. 北京：人民邮电出版社，2013.

[7] （美）维克多·迈尔–舍恩伯格. 大数据时代：生活、工作与思维的大变革[M]. 盛杨燕，周燕译. 杭州：浙江人民出版社. 2013.

[8] 王劲. 大数据时代的管理变革[J]. 中国商贸. 2013（02）：190.

[9] 宋宝香. 数据库营销：大数据时代引发的企业市场营销变革[J]. 价值工程. 2012：132–133.

[10] 许继楠. 医疗服务业率先受益于大数据[N]. 中国计算机报. 2012–02–20（017）.

[11] 苏萌. 大数据时代的商业变革[J]. 信息与电脑. 2012，11：17.

[12] 傅琳雅，傅琳晶. 大数据时代的营销革命——一场席卷全球的商业变革[J]. 中国商贸. 2013（35）：21.

第10章 大数据应用案例

随着大数据技术的日益成熟，大数据的应用也越来越广泛，人们几乎每天都能够看到大数据的一些新奇的应用，它帮助人们获得真正有用的价值。很多行业都会受到大数据分析能力的影响，其中大数据在金融行业的应用尤为突出。

10.1 大数据在金融行业中的应用案例

10.1.1 摩根大通信贷市场分析

固定收益研究与分析人员长期以来都在为客户寻找更直观与精准的数据分析与呈现方式。使用传统的分析技术例如在线图表、表格、共享文档等方式来呈现与分析庞大的信贷市场数据是极其复杂且耗时的工作。从这些海量的数据中寻找并完成特定报表是非常困难的，并且查找逻辑时往往出现大量不相关的数据。时至今日，互联网毫无疑问是查找、传输市场数据的必要途径，但要更快速有效地为客户提供价值，金融机构需要更直观的数据呈现与浏览方式。

摩根大通银行（J.P. Morgan）是摩根大通集团的子公司。作为世界顶尖金融机构，摩根大通为一百多个国家的客户提供资产管理、投资银行、商业银行、私人银行以及证券管理等服务。

像摩根大通银行这样的大型机构需要不断创新的数据分析手段来应对爆炸式增长的数据，并且要给与客户最简洁直观的数据呈现方式，二者不可或缺，才能在竞争激烈的金融市场上树立自己的品牌。

摩根大通银行利用Datawatch Panopticon创造了非常直观易懂，且能够实时更新、紧跟市场动态的“信贷市场动态图”，如图10.1所示。这个工具使得用户能在定制化的面板中，以颜色、大小、近似值等方式呈现和分析数据，帮助用户轻松地分析趋势，发现规律，作出决策。

从图10.1中可以看到“信贷市场动态图”以一组特定颜色，大大小小的方块分别代表公司债券市场的诸多基本信息。图表按不同行业划分区块，每个区块内的方块代表本行业所发的债券。方块大小代表债券的发行量，颜色代表此债券的表现。通过此图可帮助投资者一目了然地看到哪些行业，哪些债券受到追捧，以及这些债券的发行量是否满足自己的投资需

要。点击各个方块，则可看到相应债券的详细信息，如债券评级，哪些分析员负责研究此债券，以及他们的直线电话，并且附有详细研究报告的链接等。

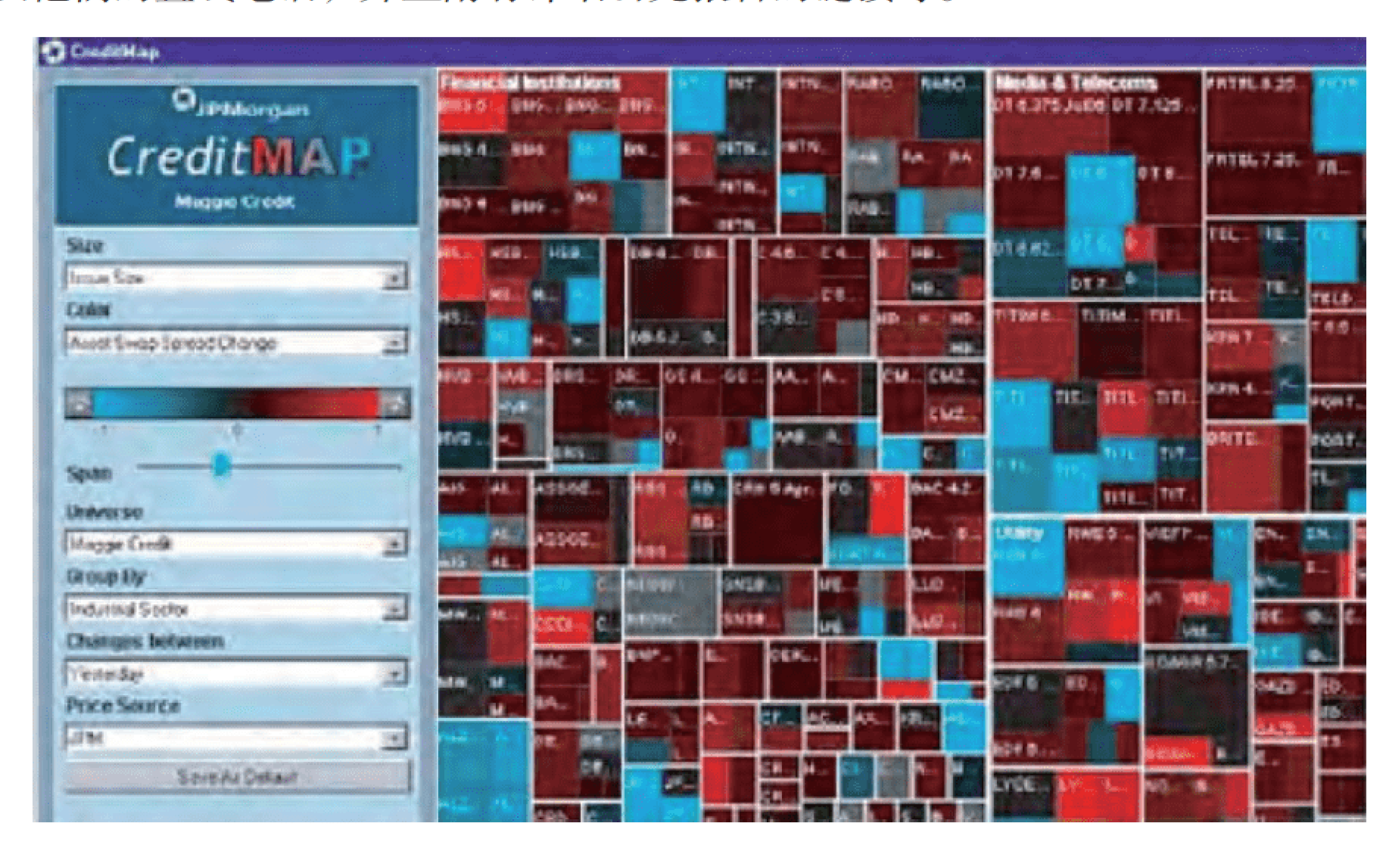


图10.1　信贷市场动态图

在采用Datawatch Panopticon之前，摩根大通银行的客户需要从文本报告的无数报表中来寻找公司的债券信息，但由于庞大的市场规模以及海量的债券数据，客户很难从大量的信息中提取最需要的部分，因而错过了不少难得的投资时机。

摩根大通银行在固定收益研究方面是市场领导者。使用了Datawatch Panopticon方案后，摩根大通银行的市场竞争力继续增强，并不断地改进现有“信贷市场动态图”，以帮助客户更快、更好地实现投资需要。

通过使用Datawatch Panopticon解决方案，摩根大通银行用最新、最直观的表现方式呈现出现有的大量的数据信息。一目了然的鸟瞰图可帮助客户快速了解最新趋势，单击单个方块可以继续深入研究单个债券的详细研究成果。此方案的使用帮助摩根大通银行获得了“最佳固定收益在线研究奖”。

摩根大通银行欧洲投资组合与指数总监Lee Mcginty表示：“Datawatch Panopticon帮助我们极大地增强了对信贷市场数据的分析能力，是我们获得这个全欧洲第一的奖项不可或缺的一部分。”

10.1.2　奥马哈外汇风险敞口和实时数据分析

奥马哈外汇要求实现实时对外汇数据流中CEP引擎中的数据进行可视化分析，并建立树形图和可视化热图，来全面取代传统的报表工具。奥马哈外汇采用的解决方案是用Datawatch连接到多个CEP引擎，在“任务控制”环境下将28个分析仪表盘显示在非常大的视频墙上。实现以客户为指导的部署，同时最小化Datawatch咨询支出。Datawatch团队在30天内完成了部署，达到了所有交易都可以使用交互式仪表盘来分析实时数据的目的，在高度变化的市场条件下保持风险敞口，如图10.2所示。

图10.2　奥马哈外汇

通过融合Datawatch的热点地图，人们实时分析风险的能力在今天这样高度不稳定的市场中将会大大提高。

10.1.3　瑞士银行集合风险分析

市场波动性的增加，意味着金融服务公司必须洞察全天候的风险，进行充分的风险分析。运用Datawatch数据可视化软件，基金管理人可以监控风险限额的使用情况，并借助外部风险引擎，分析风险数据。仪表板可以帮助用户找出问题、发现机会，它比任何传统的报告系统都更快，更高效，如图10.3所示。

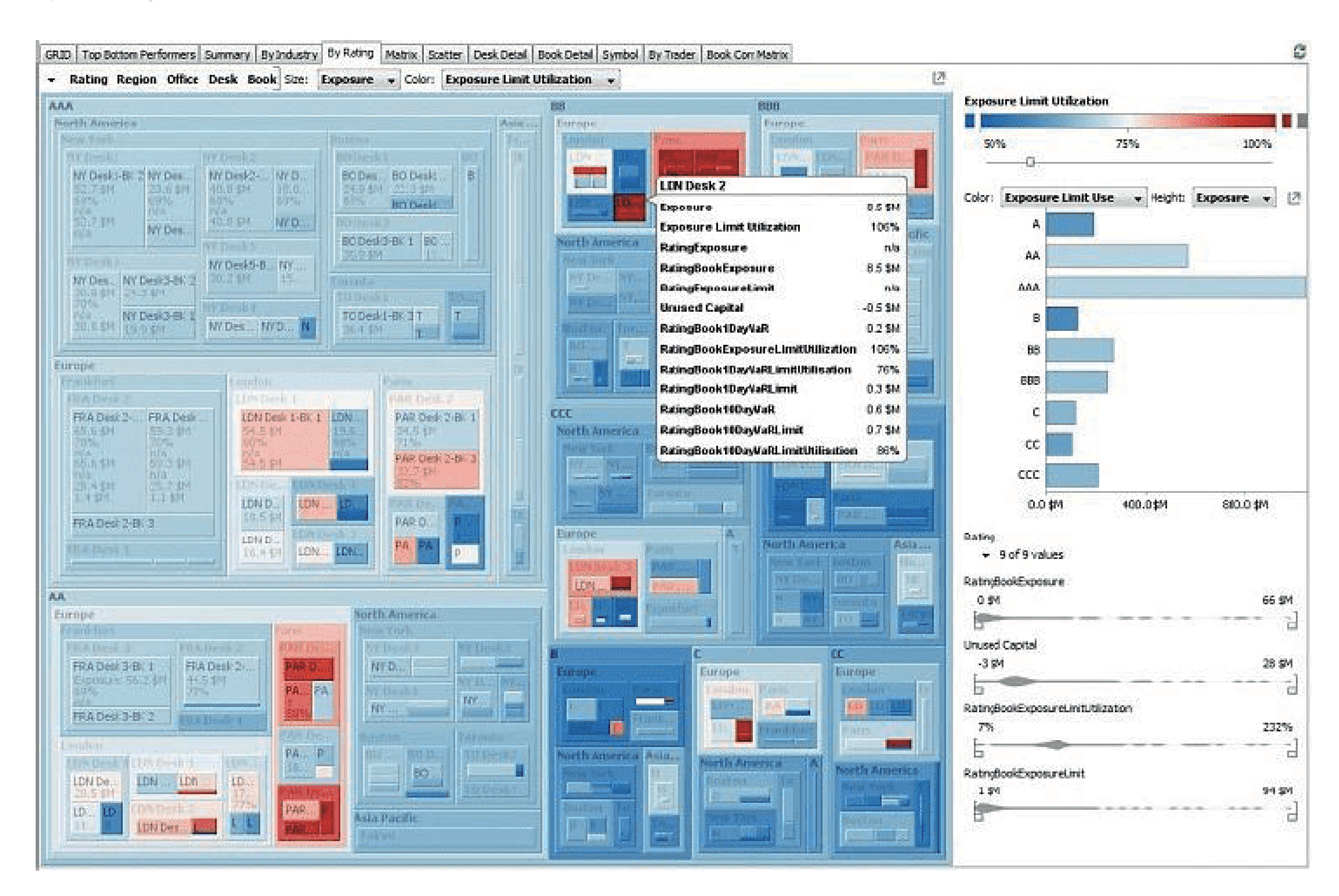


图10.3　某瑞士重要银行仪表分析盘

Datawatch作为一家金融服务公司，除了努力满足并超越客户、员工和股东的需求，还需要合适的工具，以便能够迅速、容易、经济地做出更好的业务决策。下面简单介绍Datawatch

与其他工具结合使用进行风险分析的一个案例。

该银行需要在任意时间点都能为交易员提供能显示VaR的仪表分析盘（包括实时的和历史的数据），希望对所有的办公室、个体和产品组合进行分析集合的风险并且允许各个职位的工作人员可以深度研究其中的细节，并且能在一个仪表分析盘内部直接链接到交易系统与现有的IT基础设备进行交互。了解了该银行的需求并进行分析后，Datawatch团队把Datawatch可视化嵌入到现存的优先交易系统中（Java环境），在Datawatch顾问团队支持下，由客户管理的第三方开发完成。此次实施该解决方案共历时120天，完成后提高了该银行办公的合作效率，且交易员和经理之间更容易实现结果共享；风险经理能够按需查看VaR数据，计算风险的大小，实现更高效的操作，从而进行风险分析，监控风险限额的使用情况，以及分析由外部风险所带来的风险大小。另一方面，也可以预防风险，Datawatch的实时技术帮助该银行直观地监测数据流、数据队列和交易效率。

该银行的首席风险控制官对产品的评价是“我们已经在我们的交易系统里做了一个可持续的投资并且这个系统现在很好地为我们工作。我们决定把Datawatch 工具用SDK直接嵌入我们的系统。现在我们拥有一个完全无缝的用户界面，而可视化成为我们理解我们数据的主要能力”。

10.1.4　汇丰银行多维度的历史数据分析和异常值快速分析

Datawatch具有强大的功能，除了使风险分析更有用、更高效外，其功能还包括从CEP引擎订阅完整的参数化数据流；使用时间窗口和时间段分析；找出时间段内增量变化；能够使用快速的内存OLAP数据模型监测和分析多维数据；能够可视化趋势、聚类、相关性和异常值等。例如，图10.4显示的是某主要国际银行的仪表分析盘。汇丰银行就是利用这些功能来完成多维度分析以及异常值分析的。

图10.4　某主要国际银行的仪表分析盘

该银行需要分析标记的历史数据，在可交互的仪表分析盘中，利用CEP数据，利用网络传递至用户，并嵌入到用户的桌面应用中。实现了多维数据分析，达到快速分析异常值和比较不同时间段数据的能力。该银行采用的解决方案是把Datawatch链接到Kx Kdb+ tick数据库和Sybase CEP引擎，由Datawatch的顾问支持的客制化实施。此次解决方案在30天内部署完毕，完成后用户就可以创立自己的仪表分析盘了。

投资组合主管 杰森·康勒表示："我们的分析师现在能够建立并且发布完全由他们的独特方法产生的个性化仪表分析盘并用其去分析市场数据。他们的合作效率更高了，并且能够用更短的时间去发现异常值。"

10.1.5　对冲基金选择Datawatch来观察实时的市场流数据

Datawatch被世界各地的知名银行、对冲基金和券商所广泛采用。公司高管、基金经理、交易员和分析师都频繁地使用这个先进的数据可视化工具来引导业务盈利，同时密切关注风险。

"实时"的概念是非常模糊的，因为几乎所有的软件公司都表示，他们的可视化系统能处理"实时"的需求。然而，绝大多数情况下，他们的意思是，每当有新的数据需求时，其软件可以从外界数据源获取更新。比如，某系统可能发出更新需求，以便产生新的图表。该系统从外界资源获取更新数据，并且在获取数据的同时展示在显示器上。在特定的时刻，这些数据是精确的；然而精确性立即就会过期，因为程序在执行完整个刷新过程之前，是不会进行可用数据升级的。

Datawatch可以实现真正的实时分析，像CEP引擎和消息总线这样的外界系统将不断把数据推进到系统中——没有定期间隔，而是即刻不间断地分笔进行。比如某对冲基金要求不包括内部IT的小型公司，做成可视化实时的市场流数据，实行快速部署。Datawatch团队进行需求分析后，决定由交易员自身实施解决方案，把Datawatch链接到Thomson Reuters 和 Bloomberg feeds 产生的数据。解决方案实施后，该对冲基金能进行更加有利可图的交易，用更少的时间搜索交易，更容易地管理多个战略。图10.5显示的是某对冲基金的仪表分析盘。

图10.5　对冲基金的仪表分析盘

康勒 信息部副总表示：“Datawatch 让我可以更便利地跟踪个别股票和不同分区。Datawatch这个可视化工具包已经在交易方面对我产生了极大的帮助。”

10.1.6 衍生品交易公司的交易活动的浏览与分析

某衍生品交易公司要求识别期权定价异常，检测订单活动和模型定价异常，进行时间序列数据分析，进行可交互树图可视化，并链接到多个SQL数据库。Datawatch团队对这些需求分析后，决定由客户来主导实施。Datawatch被部署到5个仪表盘设计师和37个交易员的终端系统上，客户团队开发了六个不同的仪表分析盘（如图10.6所示），用于比较按市值计算和按模型计算的不同交易结果。此次解决方案用7周完成了部署，所有的交易员和分析师都拥有了客制化的仪表分析盘，在单个仪表分析盘中首席投资官能够浏览全公司业务范围的交易活动。

某衍生品交易公司首席投资官表示：“我们需要一个可以支持交易员并且为高级管理者提供综合分析视图的平台。我们也需要一个优化金融时间序列数据的系统。很显然，atawatch是一个很好的选择。”

图10.6 某衍生品交易公司仪表分析盘

10.1.7 跨国保险公司连接多个数据库来进行风险分析

Datawatch几乎可以连接任何数据源，包括即时数据串流、消息总线、复杂事件处理（CEP）引擎的输出、OData源、平面文件、关系数据库和专有数据格式等。某跨国保险公司应用Datawatch提供的功能来实现其风险分析。

该保险公司需要对多个产品组合进行实时风险分析，连接到CEP引擎和相关数据库，通过网页部署、分析CEP引擎的多维数据流，实行快速增长的数据设置。该解决方案由Datawatch团队主导实施，将Datawatch连接至CEP引擎数据流、Oracle数据库、SQL Server 数据库和Thomson Reuters高速分析数据库。Datawatch团队在30天内完成了部署，平台能够支持未来至少五年的数据要求，所有分析师和交易员的桌面都部署了交互式仪表分析盘，如图10.7所示。

图10.7　某跨国保险公司的仪表分析盘

某跨国保险公司首席信息官表示：“我们选择Datawatch EX作为风控系统前端，因为它具有实时分析显示的能力，大范围的可视化和公司成熟的金融应用程序。”

10.2　大数据在医疗行业中的应用案例

早在很久以前，医疗行业就遭遇着海量数据、半结构化和非结构化数据的挑战与机遇。而大数据的出现为海量医疗数据的分析挖掘提供了可能。新一代医学技术的出现，也标注着医学研究已经进入大数据时代。在此，笔者介绍一下目前大数据在医疗行业中的几个应用案例。

10.2.1　美国糖尿病患者分布情况分析

由美国疾病预防和控制中心（CDC）来提供数据，现在Datawatch可以进行糖尿病患者分布情况的分析。将Datawatch可视化系统嵌入到现有的数据系统中，在单个界面中就能够浏览全美国糖尿病患者的分布情况（如图10.8所示）并能对比各个州的患病率。用简单直观的图形和颜色来对比美国糖尿病患者绝对数量及相对数量的分布情况（如图10.9所示），从而分析出影响糖尿病的各种地理环境因素。

考虑到糖尿病涉及面广、复杂和多维的特点，通过对大量糖尿病患者的详细资料，如就医情况、健康数据、病情史等进行分析，对比糖尿病的发病率与贫穷率、人口、家庭收入、教育程度之间的关系，从而揭示出可能对健康产生影响的社会因素和环境因素，如图10.10所示。

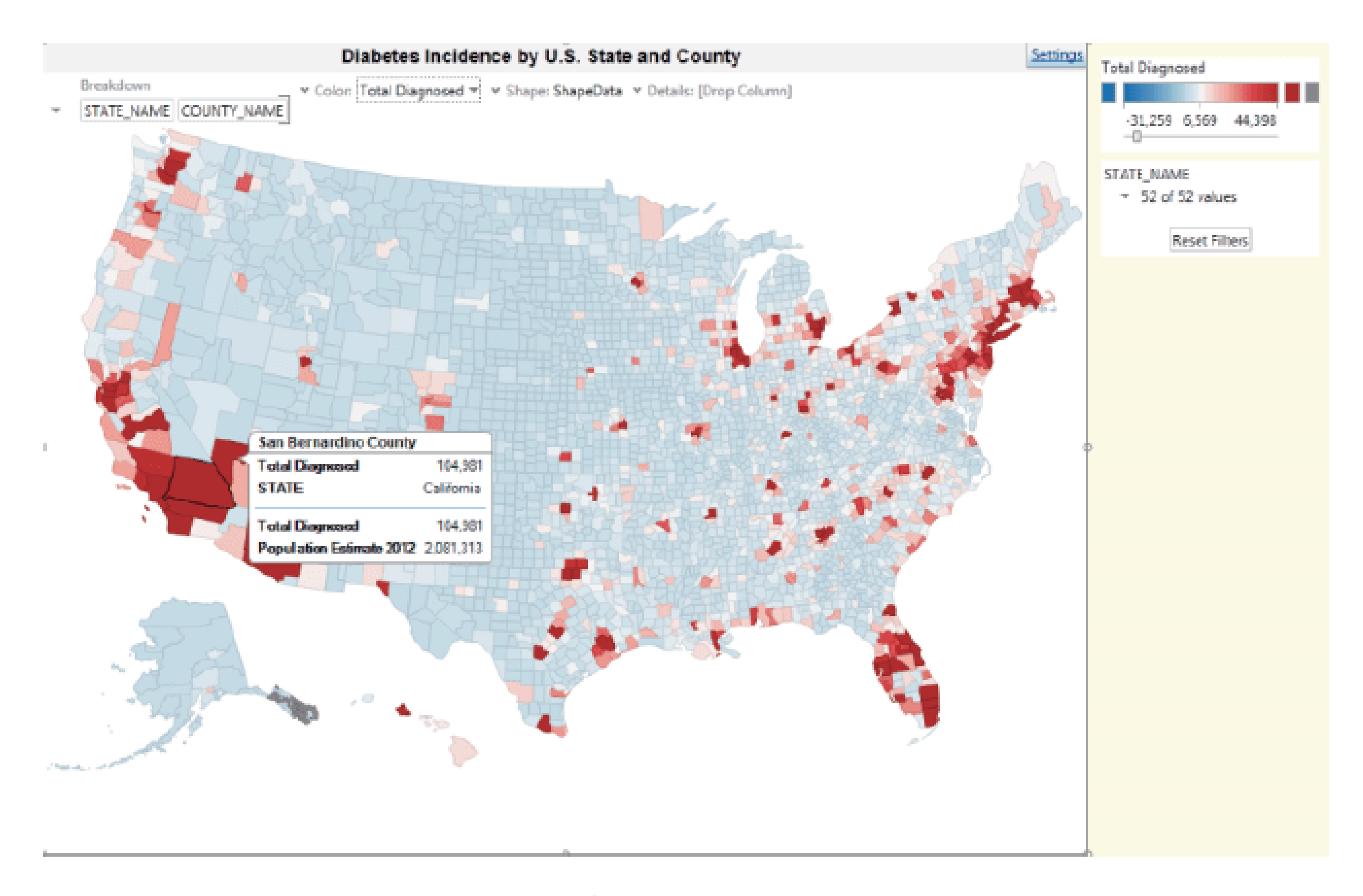


图10.8　糖尿病患者在美国各州的分布

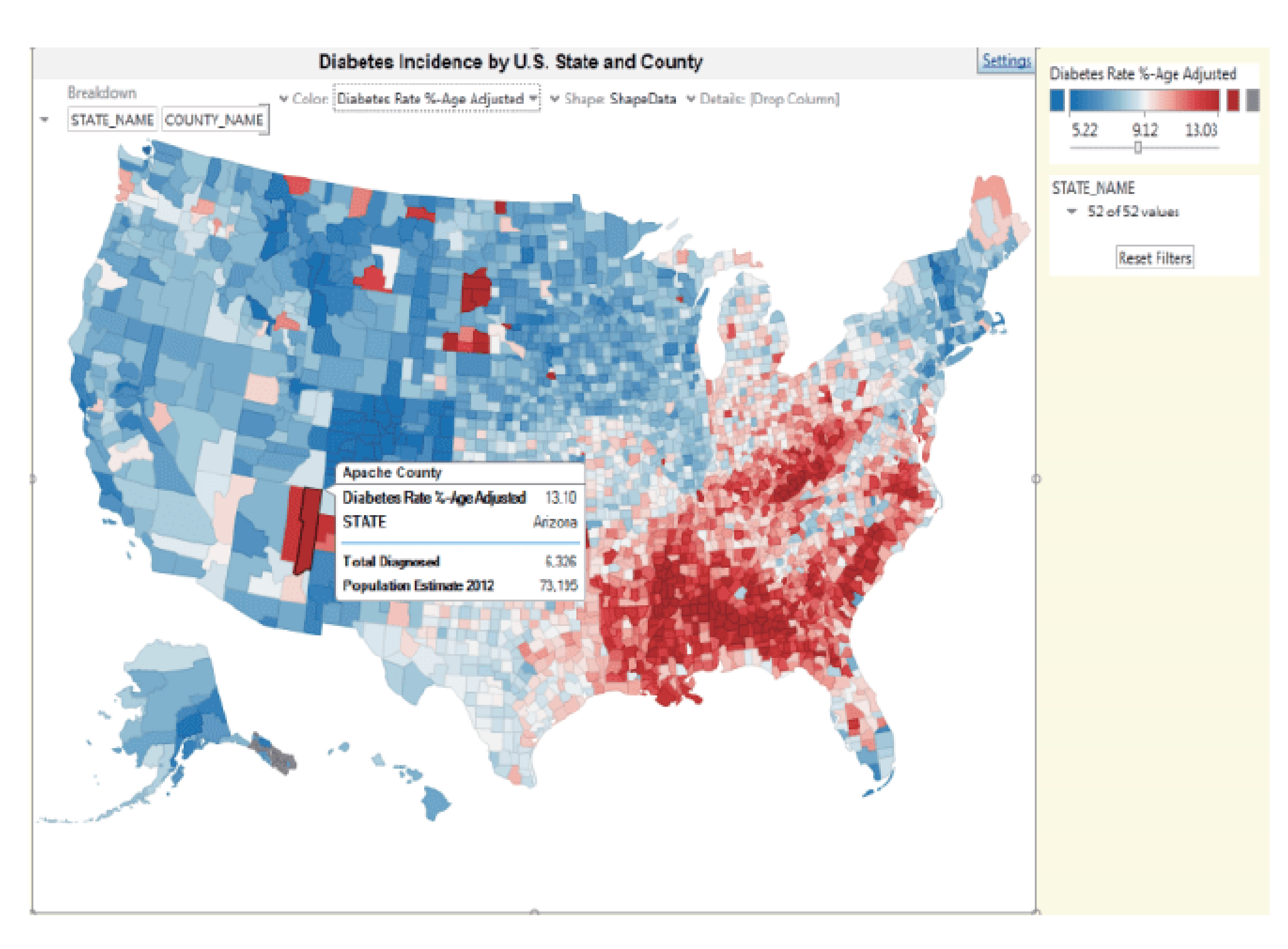


图10.9　糖尿病患者各年龄层在美国各州的分布

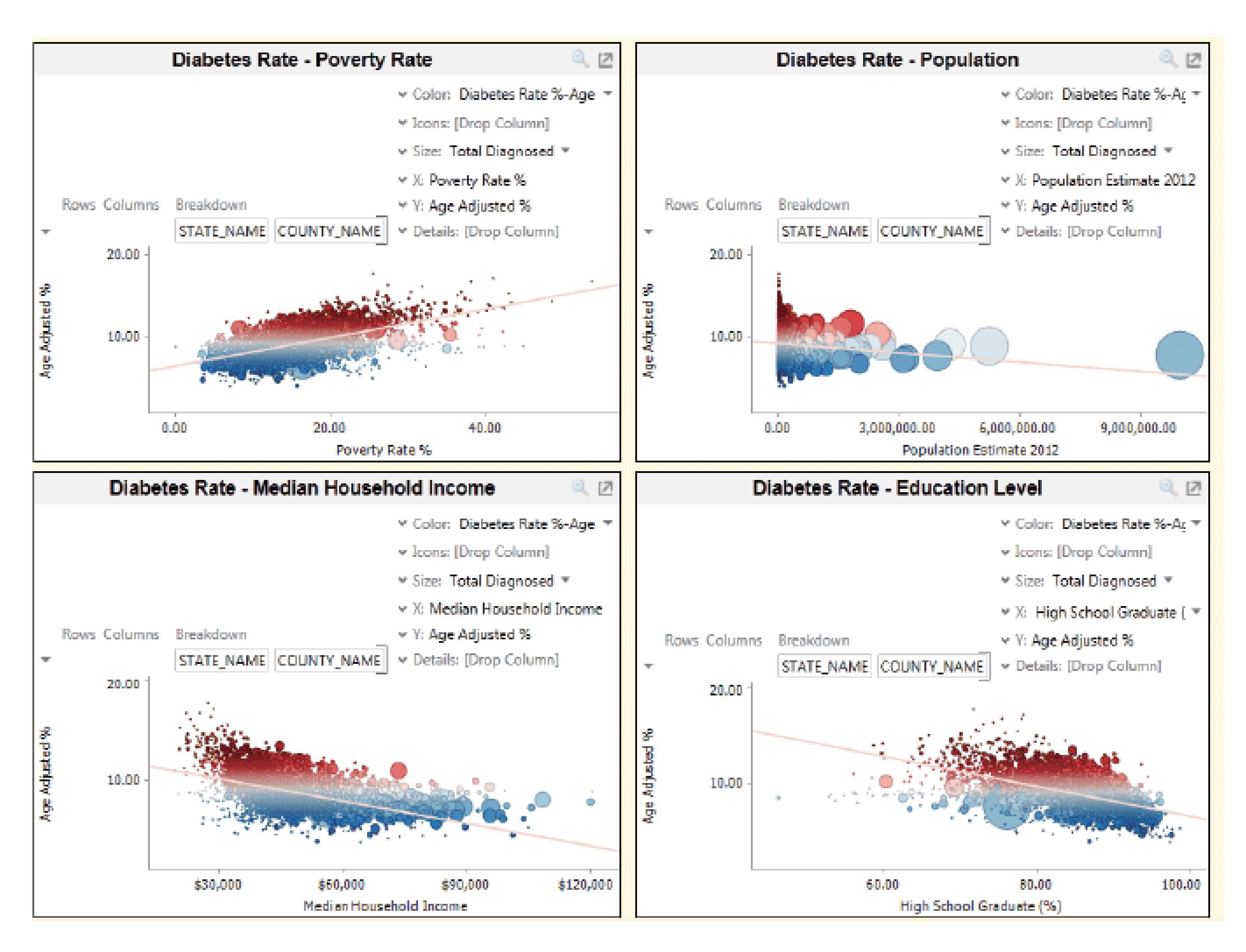


图10.10　糖尿病与贫穷、人口、家庭收入以及教育程度的关系

10.2.2　医疗机构病房的实时监控

在传统的模式里，患者的病例数据是以文本或者表格的形式显现出来的，繁琐而复杂，十分不利于隐藏信息的捕捉和获取。所以，医疗机构要想达到对病房实时监控的目的，就显得十分困难了。

Datawatch的出现能够扭转乾坤，它实现了在一个仪表分析盘中直接链接各个病房并与现有的医疗信息基础建设相融合，让医疗机构可以更加便利地跟踪每位病房患者的动态情况，实现了对于病房患者的实时监控，如图10.11所示。接下来就Datawatch在医疗机构病房实时监控方面的强大功能来做一个比较详细地说明。

图10.11　实时监控患者体征动态一览图

图10.11显示的是病房实时监控下患者体征的动态变化情况，从这里能够看到病房不断波动变化着的动态效果，但是此图只能以某一个时间点的图示为例来说明（MICU，medical intensive care unit，内科重症监护室；RR，relative risk，相对危险度；LFHF ratio，low frequency/ high frequency，高低频之比；TP，total protein，总蛋白）。

从图10.11中可以非常清晰直观地看到房间的整体布局，同时还能获知病房患者的分布情况。比如哪些房间没有患者，哪些房间有患者。圆圈的大小表示高低频之比（LF/HF）的大小（这是衡量心脏功能的一个指标），圆圈越大，表明比值越大，反之，比值越小；圆圈的颜色表示患者相对危险度（RR，这里特指发生医院感染的相对危险度）的高低，红色表示相对危险程度高，最小的是蓝色，其次还有绿色、黄色等。图形颜色以及大小的动态变化非常直观，医护人员能够及时地依据这些指标做出判断，从而给患者提供更好的服务。此外，还可以从该图右上角看到病房患者的年龄、性别的分布情况，不同的颜色表示不同的年龄段，区域面积的大小表示患者的多少。从中可以很直观地判断出男性患者多于女性患者，并且男性患者中以41～59岁年龄段最多，而女性患者以61～79岁年龄段最多。另外，还可以从不同的角度去直观地分析此图，例如当我们做出不同的选择时，可以很清晰地认识到不同的颜色还可以表示不同的年龄段以及不同的相关病史等，如图10.12所示。

此外，当医生想了解某一位病人的一些具体情况（例如病人的通气状况、相关病史、诊断手术情况等）或者相关指标的具体值时，通过简单的操作就可以获取（见图10.13）相关信

息，这有助于医护人员快速有效地依据综合因素制定诊疗计划，极大地提高了医疗效率。而这对于患者而言是至关重要的。图10.14所示为病房实时监控呈现的患者相对危险度。

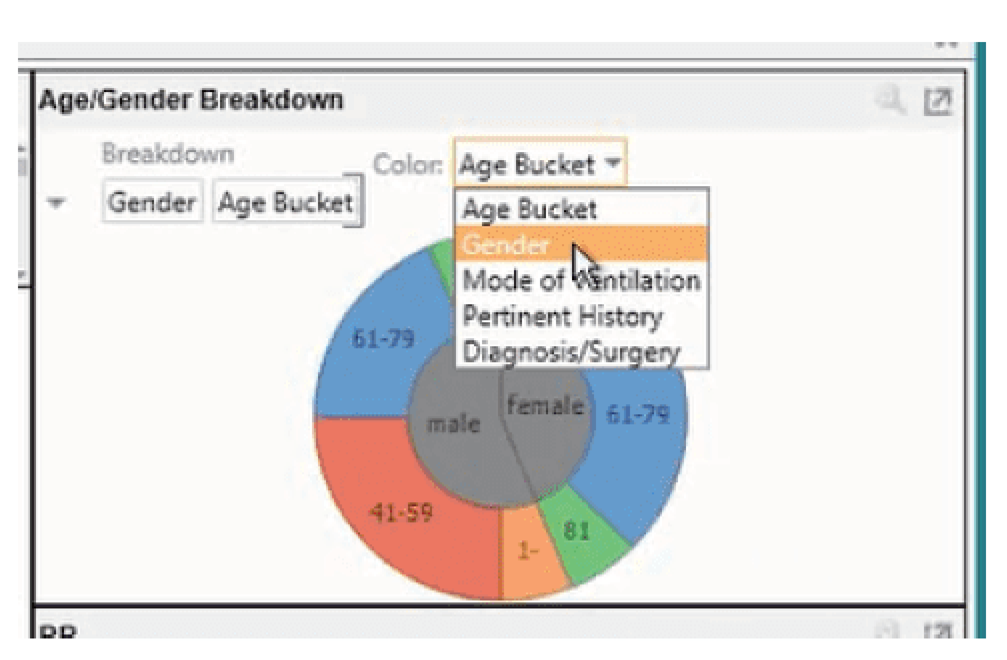


图10.12　病房患者年龄/性别呈现

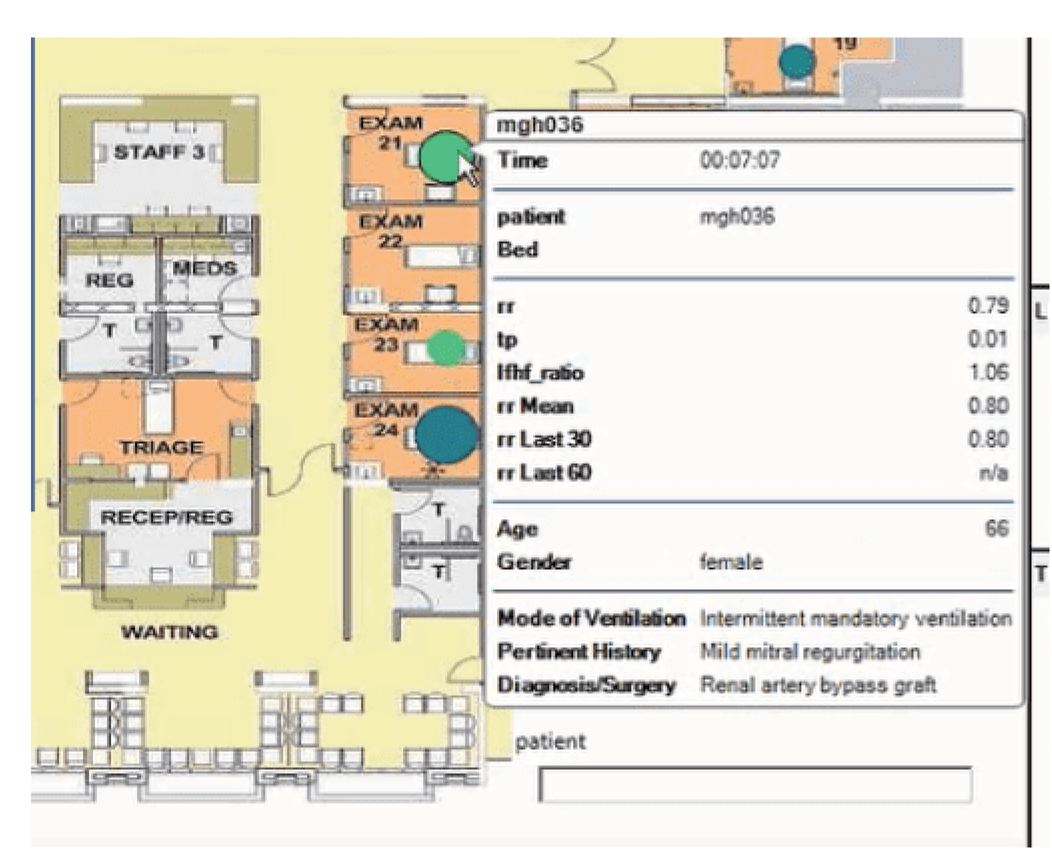


图10.13　病房患者信息的具体呈现

Datawatch可以充分利用点状图、树形图以及线形图把病房患者指标的动态变化以清晰直观的形式呈现出来，如图10.14所示。在左侧的点状图中，颜色代表患者的相对危险度（RR），红色表示患者的相对危险度高，蓝色表示患者的相对危险度最低，其次还有绿色、黄色，当然，最引起医护人员注意和重视的无疑是“红色”的出现。除此之外，圆圈的大小代表总蛋白（TP）的多少，当圆圈越来越小时，预示着患者总蛋白的急剧下降，这时就会引起医护人员的注意并及时地采取相应的解救措施。树形图中方块的大小代表总蛋白（TP）的多少，方块的颜色表示相对危险度（RR），其具体的意义与点状图是一致的。该图右侧同样也是综合利用线形图、树形图以及数据的直接变化充分呈现出相对危险度和总蛋白两个指标的实时变化。Datawatch这种丰富的可视化视图给人们带来耳目一新的感觉，清晰直观的画面感特别符合人们对颜色、大小变化的这种高敏感度，有助于医护人员及时地作出判断。当然，除了上述介绍到的有关患者的某些生命指标以及相关情况，Datawatch可以根据医疗机构的具体要求做出相应的实时监控，以满足实际需求。

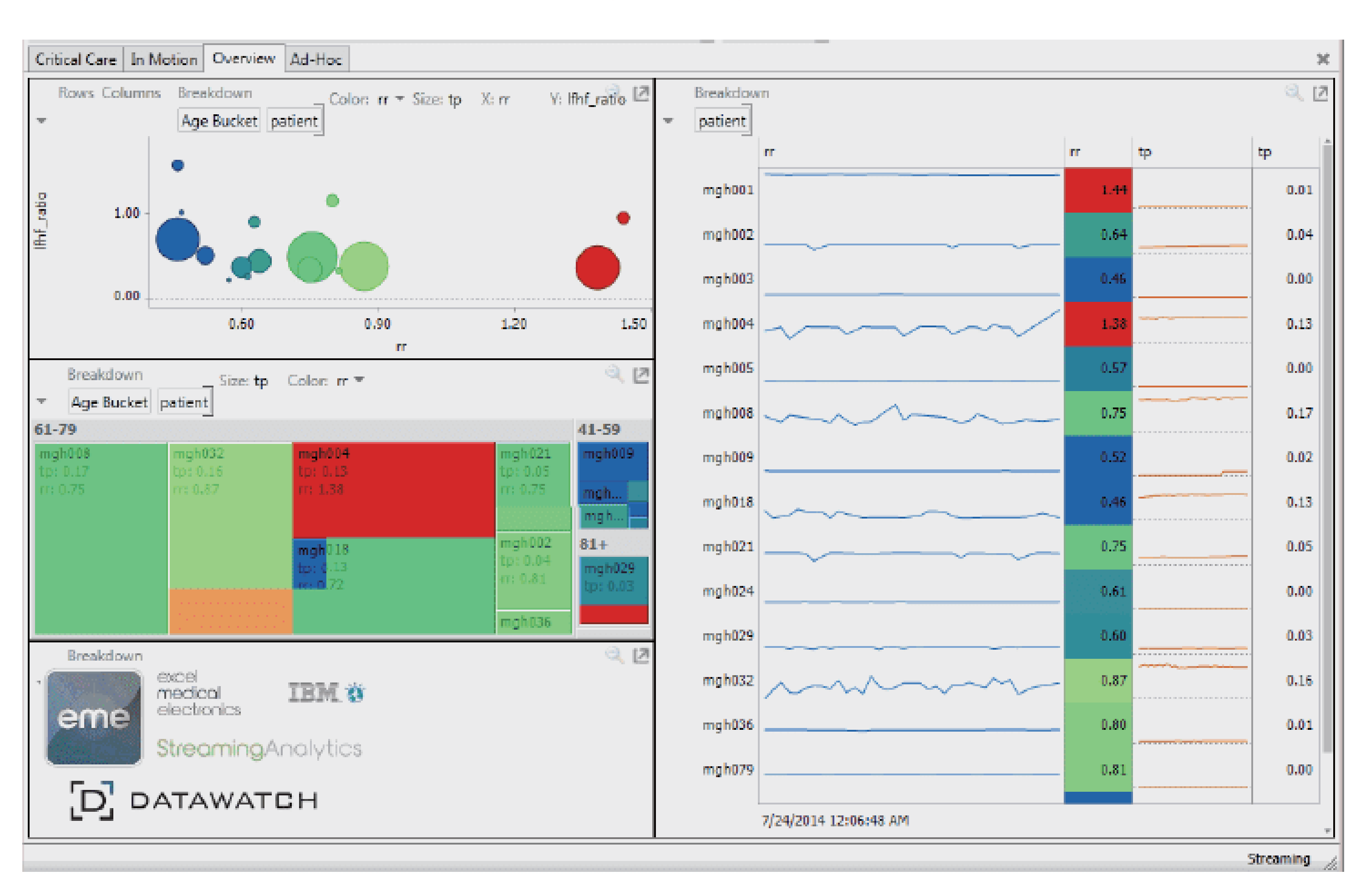


图10.14　病房实时监控呈现的患者相对危险度

10.2.3　流行病学研究

在此本书引用了复旦大学《上海地区门诊儿童流行性感冒的流行病学研究》这篇硕士毕业论文的实验数据，通过采用Datawatch的可视化技术，对论文的实验结果进行了多维度的直观展示，阐明Datawatch工具在流行病学研究上可以提供的帮助。

首先展示的是Datawatch采用热力图、树形图和饼状图三种不同图形多维度分析儿童的临床特征，如图10.15所示。

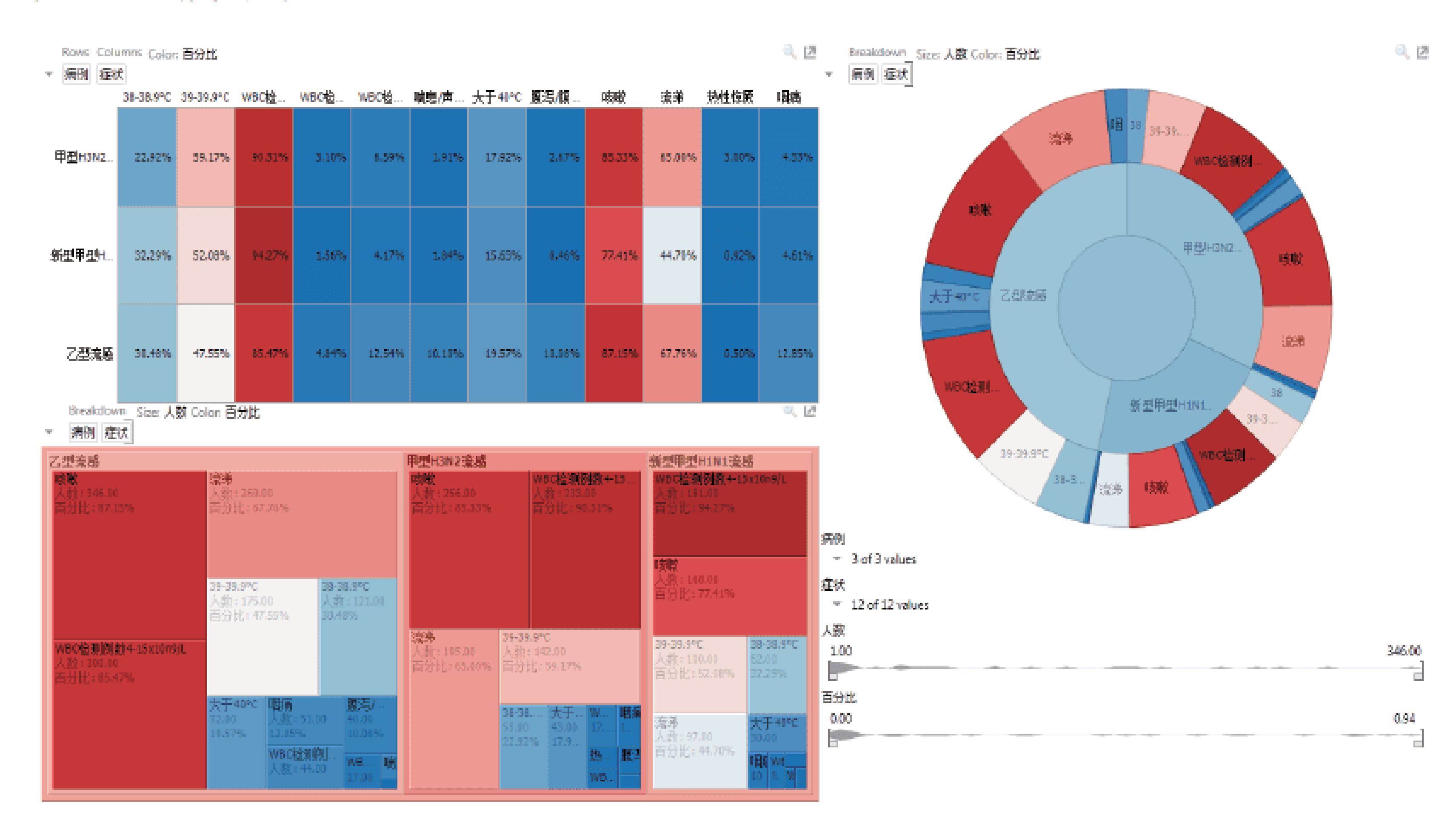


图10.15　三种流感的临床特征对比分析

Datawatch工具自带了多种图形，相同的实验数据可以用不同的图形来展示，不同的图形侧重点不一样，研究者可以根据实际需要选择最合适的展现方式，或者采用两种或多种图形，相互辅助诠释数据信息。

例如，将图10.15中的热力图放大，如图10.16所示。其中，颜色代表百分比，同列中颜色越深代表在该流感中出现这种临床症状的百分比越高。该热力图针对同一临床特征在三种类型的流感出现的百分比作出了清晰直观的呈现，简洁明了，通俗易懂。

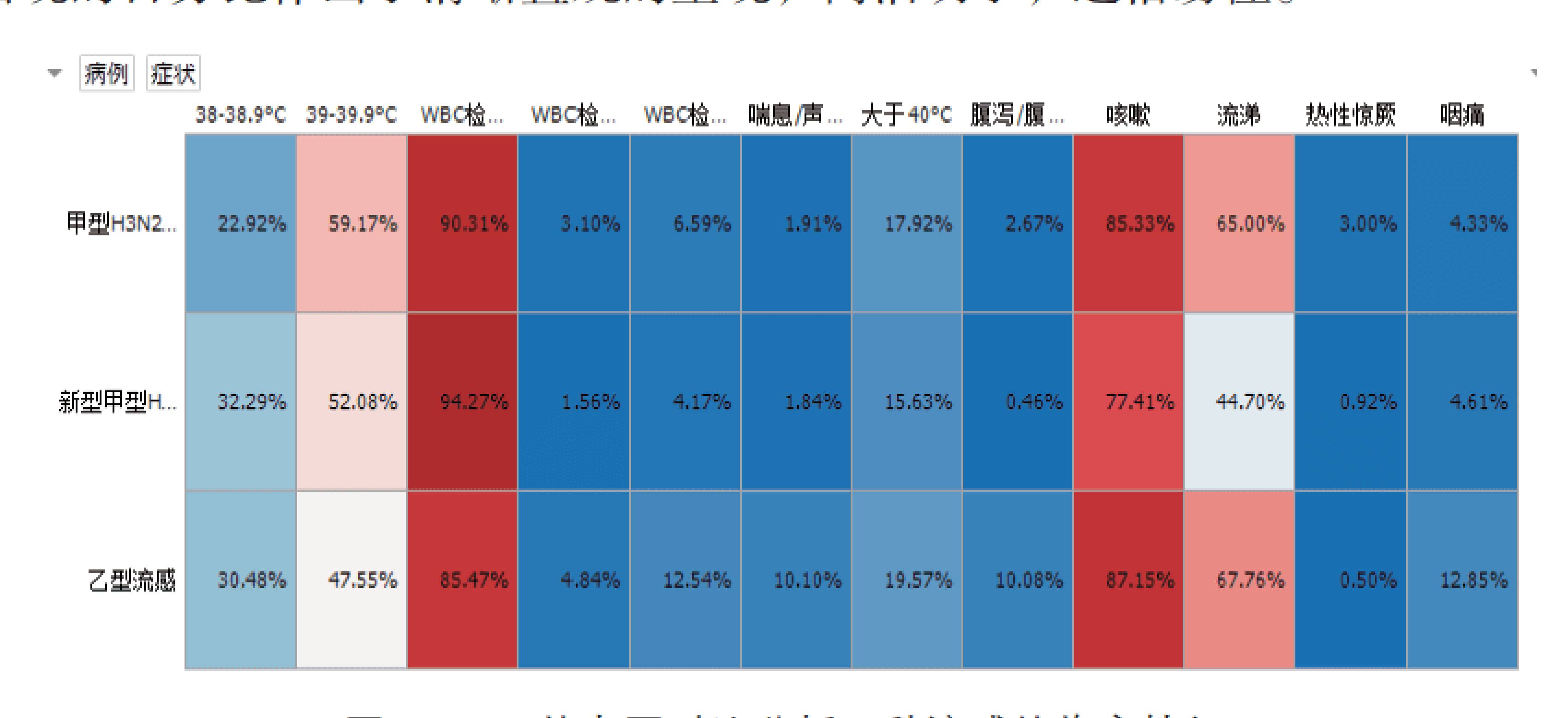


图10.16　热力图对比分析三种流感的临床特征

接着采用柱状图、热力图以及饼状图进一步分析里面的WBC监测指数及热风指数。如图10.17所示，其中柱状图的颜色代表百分比，柱高代表人数；热力图颜色代表人数，该图呈现了三种不同类型的流感在不同WBC范围内的患者数。饼状图尺寸代表人数，颜色代表百分

比。这里Datawatch采用三种不同的视图形式对同一数据进行了可视化分析，能让用户十分直观地了解输出的信息。

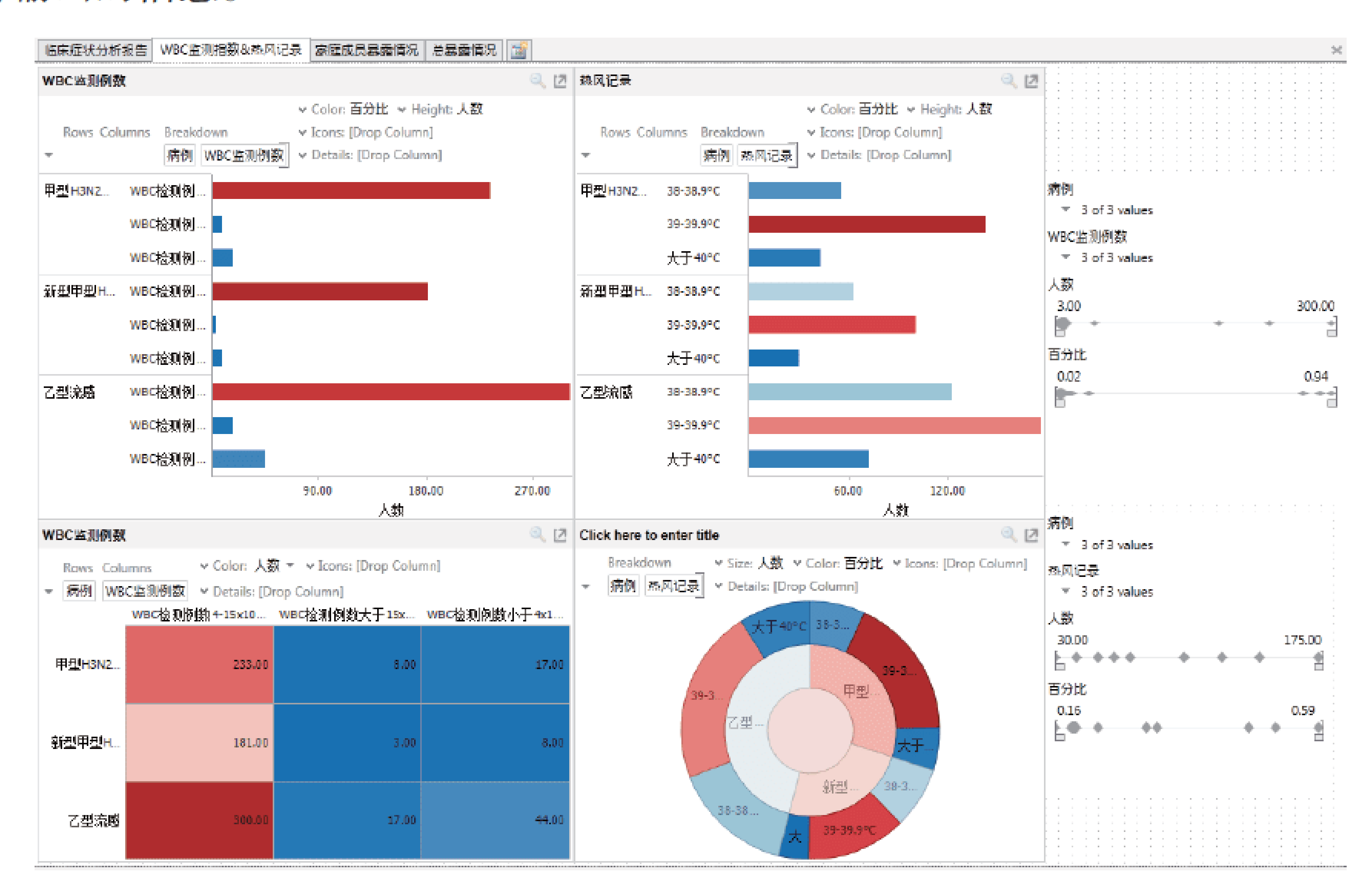


图10.17　三种流感的WBC监测指数及热风指数对比分析

总体上，从三种流感的各类临床特征图像的对比分析不难看出，甲型H3N2伴随流涕和咳嗽更常见，出现热性惊厥的比例更高；乙型流感伴随喘息/声嘶的比例、咽痛以及腹痛症状高于其他两种；H1N1患儿出现腹部症状及咽痛的比例较低；WBC检测显示绝大部分儿童符合病毒感染指标等。

此外，研究者发现儿童患者在发病前，有的明确与发热或急性呼吸道感染患者密切接触史，为此对季节性H3N2、新型甲型H1N1和乙型流感的暴露对象情况作了分析。分别采用树形图和点状图进行展示，如图10.18所示。

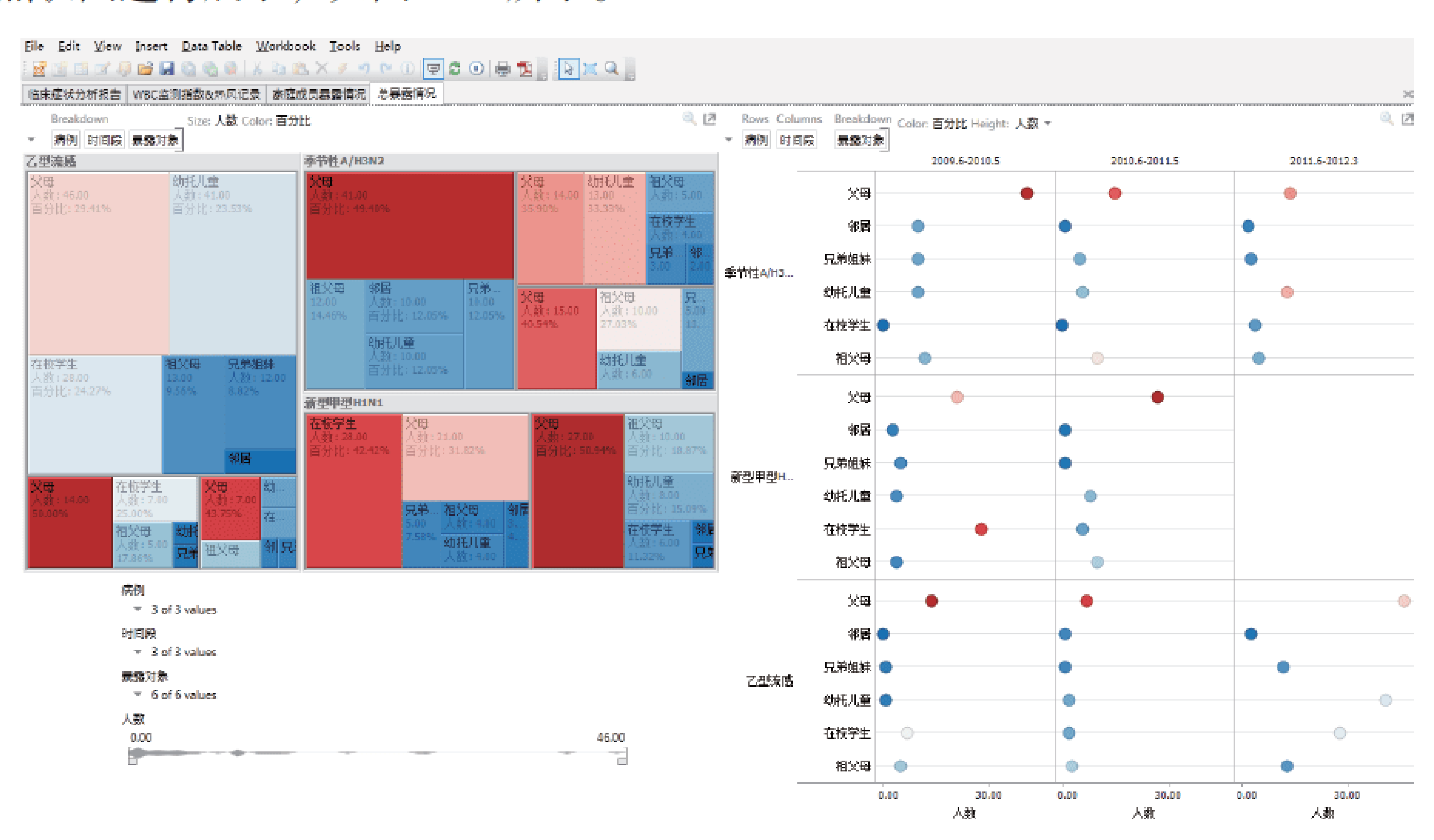


图10.18　三种流感暴露对象情况分析

树形图的方块大小代表人数，颜色代表暴露对象百分比。点状图中颜色同样代表暴露对象百分比，Y轴表示暴露对象类别，x轴表示暴露对象数量。这样可以很直观地看出，三种亚型暴露对象均以家庭成员为主，占45%～81%，其次是在校生或幼托儿童，如图10.19所示。

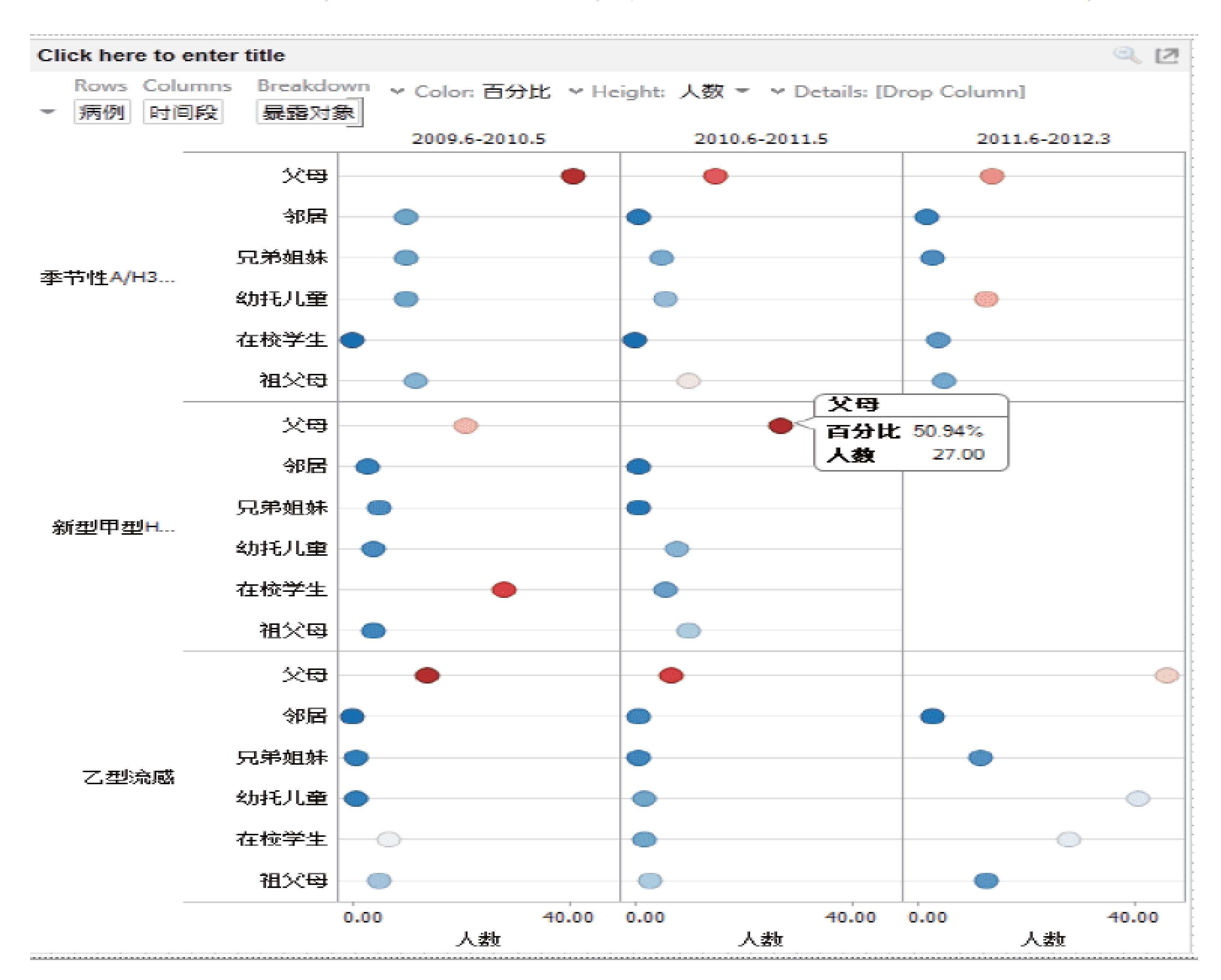


图10.19　点状图分析三种流感暴露对象情况

接着进一步对家庭成员的暴露情况进行分析，如图10.20所示，块的大小代表人数，颜色代表百分比，红色越深代表所占比例越高，蓝色则反之。从图中可以轻易判断出3种亚型的家庭成员暴露对象中父母占绝大部分。通过对总暴露情况和家庭成员暴露情况的分析，显示在家庭和集体机构中传播是流感的主要传播方式。

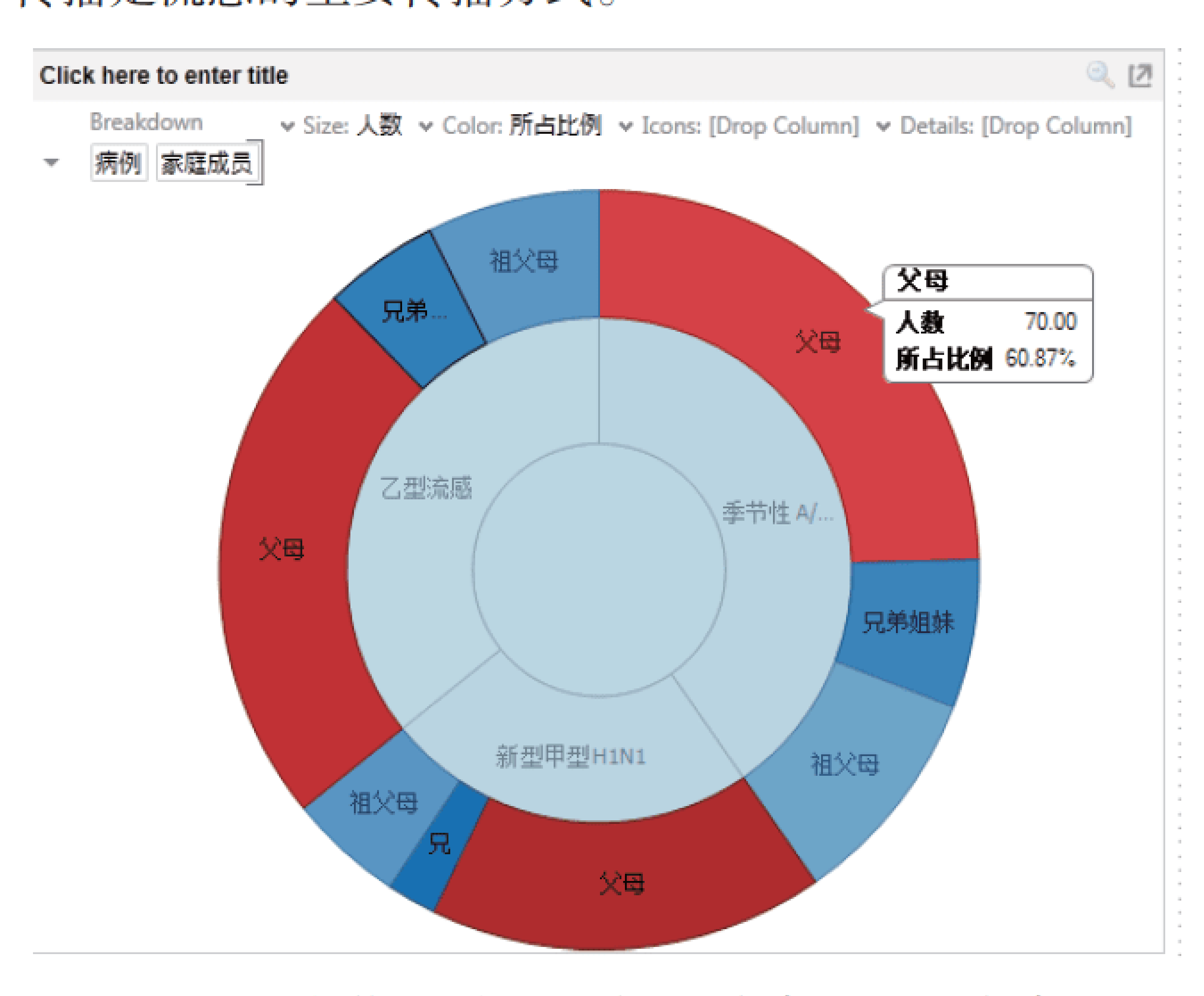


图10.20　饼状图分析三种流感家庭成员暴露对象情况

在整个流感案例里面，运用Datawatch工具，采用热力图、树形图、点状图和饼状图对研究数据进行分析。从简洁鲜明的图像中，可以更便捷地分析流感的临床特征及暴露对象情况。对不同的变量进行对比分析，帮助人们更快地还原数据背后的信息，不得不说Datawatch是医学研究的好工具、好帮手。

10.3 大数据在互联网企业中的应用案例

21世纪是信息化的世纪，“无处不在”的互联网使由此产生的用户数据越来越庞大，而随着大数据技术的日益成熟，这庞大的数据成为一笔不可估量的财富，互联网企业通过消费者浏览、购物等相关数据，来对其消费习惯、生活方式等加以分析，从而进行精准营销。亚马逊、淘宝网、Facebook等互联网企业都是善于利用大数据技术的代表企业。

10.3.1 亚马逊

亚马逊以企业云平台闻名于世，虽然该公司拥有高质量的用户数据资源和巨大的网站流量，但在之前相当长的一段时间内，亚马逊还是将主要的精力集中在产品销售上，广告只是作为其销售业务的补充。然而仅2013年一年，亚马逊的广告收入就激增了45.51%，而且仍然在持续地增长。虽然广告这项收入在集团总收入中所占比例很低，跟其他广告巨头相比也还有很大的差距，但越来越多基于海量用户数据的丰富广告产品的出现和广告带来的收入增长，正是大数据分析在企业内部应用的价值体现。

亚马逊希望通过对客户行为的分析来提升广告的投资回报率，实现精准营销。亚马逊拥有世界级的个性化推荐技术、无可比拟的用户数据和互动视频内容。目前，亚马逊正在利用这些优势，努力构建新一代的广告产品。这种新一代的广告产品就是基于强大的需求方平台（DSP）的实时竞价（RTB）广告模式，它可以实现真正基于用户兴趣的广告呈现方式。RTB模式通过对庞大用户信息的掌握和精准定位，使亚马逊能够在海量用户中将不同特点或个性化特征有差异的用户进行分类，然后将分类好的用户群与不同类型的广告进行匹配。如果广告投放的商品恰好是用户需要或者感兴趣的产品，那么广告就能产生最大的效益。

为了达到该目的，亚马逊采用自身的基础设施云服务，后台采用Hadoop技术架构的100个节点的按需弹性MapReduce集群。亚马逊Web服务的弹性MapReduce在亚马逊弹性计算云（Amazon EC2）和亚马逊简单存储服务（Amazon S3）上运行，具有极强的拓展性。如此一来，可以将处理时间从2天降到8小时，广告投资回报率也增加了500%。

2013年12月，亚马逊获得了“预测性物流”专利。该专利将大数据应用系统与物流系统相结合，使该公司能在客户点击“购买”之前就开始递送商品。该专利会根据某一特定地区的客户过往的订单和其他相关因素，预测客户可能购买但还未订购的商品，并且开始对这些商品进行包装和递送。这些预寄送的商品在客户下单之前，会先存放在快递公司的寄送中心或者卡车上，在顾客下单购买后就可以更快地送达，有效地减少了交货时间。具体流程如图10.21所示。

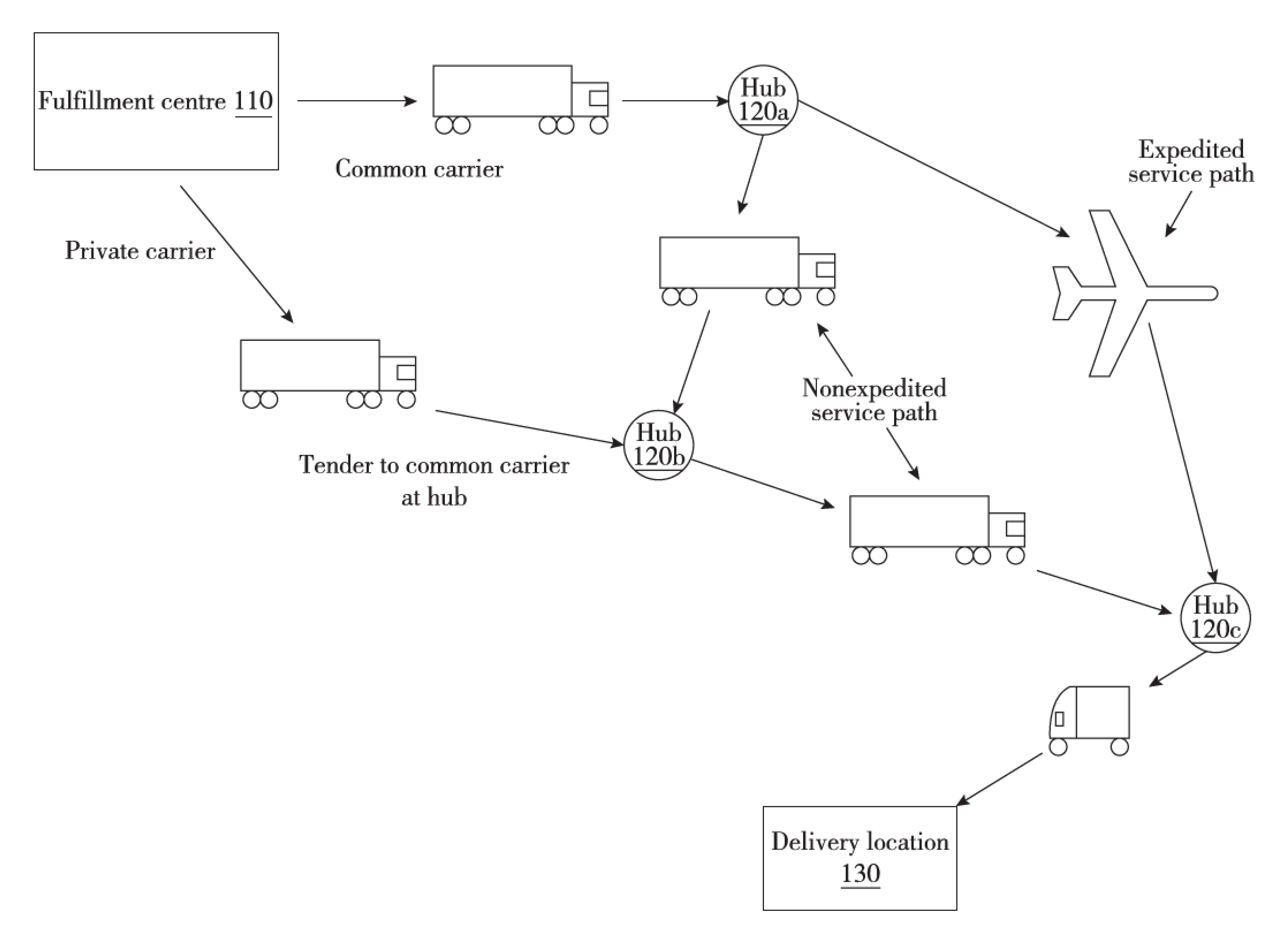


图10.21　“预测性物流”运营流程图

在对“预测性物流”商品进行预测时，亚马逊会综合考虑顾客以前的订单、搜索的产品、愿望清单、购物车的内容、退换货的情况，甚至是顾客的鼠标停留在某件商品的时长等各种因素，以提高预测的准确度。

亚马逊表示，预测性物流的方法尤其适用于预售的商品，特别是预先定于某一天开售的热门书籍等。这项专利体现了一个正在兴起的趋势——智能预测，即技术和消费企业越来越倾向于在消费者采取购买行动之前预测其需求，而不是在消费者购买之后做统计。

10.3.2　淘宝网

对于大多数系统来说，大规模小文件的读写操作是一个棘手的问题。因为大规模小文件在读取时，磁头需要进行多次的寻道和换道，导致较长的时延。在大量高并发访问量的情况下，时延将会变得非常严重。淘宝也面临大规模小图片的存储和读取这一难题。

整个淘宝网流量中，图片的访问流量就高达90%以上，而且网站中的图片数量以每年2倍的速度增长。淘宝网整个图片存储系统的容量为1800TB（1.8PB），其中被占用的空间就达到一半以上（约1PB）。淘宝网需要保存的图片总数接近300亿个，这些图片文件包括原图及其略缩图（一个原图可能要生成20个不同尺寸的略缩图）。图片的平均大小约为17KB，其中8KB以下的图片就占了60%以上。如此大规模的小图片读写给淘宝网的系统带来了巨大的压力。

为了解决这一难题，淘宝文件系统TFS（Taobao File System）应运而生，于2007年正式

上线运营。该文件系统在生产环境中应用的规模达到了200台PC Server（146G×6 SAS 15K Raid5），文件数量达到上亿级别，系统部署存储容量为140 TB，实际使用存储容量为50TB，单台支持随机IOPS 200+。到2009年6月，TFS 1.3版本上线，集群规模有了很大拓展，整个系统从原有200台PC Server扩增至440台PC Server（300G×12 SAS 15K RPM）+30台PC Server（600G×12 SAS 15K RPM），支持的文件数量也扩容至百亿级别，系统部署存储容量为1800TB（1.8PB），实际存储容量为995TB，单台Data Server支持随机IOPS 900+。

10.3.3 Facebook

Facebook是个大型社交网站，该网站每天都会产生海量的数据，其系统的数据量是15 PB（压缩以后为2.5PB），每天增加的数据量是60 TB（压缩以后是10 TB）。

在Facebook的数据分析系统中，处在系统边缘的关系型数据库系统负责进行OLTP（On-Line Transaction Processing）类的事务处理。真正的复杂深度分析，则依靠高度可拓展的Hive-Hadoop集群系统来完成。核心生产用Hive-Hadoop集群系统（production Hive-Hadoop cluster）定时导入交易数据，并完成重要的分析操作。经过分析和聚合之后的数据，可以重新载入到关系型数据库系统中（包括Oracle RAC，federated MySQL等），接受用户的查询。

为了减轻即席查询对核心Hive系统造成的压力，先将数据复制到一个备份的Hive-Hadoop集群系统（ad hoc Hive-Hadoop cluster），在此备份系统中对用户的即席查询进行处理，这样就可以隔离未经优化的查询可能给核心Hive系统造成的性能冲击，从而保证核心数据分析系统的性能。

由于MapReduce技术具有良好的可拓展性，这个系统可以使庞大的历史数据在线，对历史久远的数据也可以很快地进行分析，还可以结合新数据和新算法，对新知识的发现十分有利。

10.4 大数据在影视行业中的应用案例

提起大数据在影视行业中的应用，非常典型的一个案例就是Netflix公司利用大数据预测分析，并根据预测结果投资一亿美元拍摄了新版《纸牌屋》，该剧红遍北美风靡全球，创收视率新高，开创了大数据在影视行业的先河。除此之外，大数据还可用以分析节目收视特征、用户喜好、电影票房等，下面来看两个具体的案例。

10.4.1 大数据分析节目收视特征和用户喜好

自从2013年湖南卫视国产综艺节目《爸爸去哪儿》第一季播出以后，全国刮起了一阵星爸萌娃潮。沿袭这阵风潮，2014年夏天，《爸爸去哪儿》第二季自首播以来，反响十分热烈。国双数据中心基于自主创新的大数据视频分析平台，结合湖南卫视芒果TV的用户观看数据以及微博的相关热议，揭示了针对该剧的节目收视特征和用户的偏好。

从第二季第二期的用户24小时播放数据中可以看出（如图10.22所示），中午12点左右有一个显著的收视午高峰，下午直至晚上十点左右都有持续的播放。从播放时长来看，平均每次播放时长将近1小时。值得关注的是，凌晨1点左右有一次小的波峰，这表明有一批错过了

节目直播的忠实粉丝会通过芒果TV官网第一时间抢看点播视频。

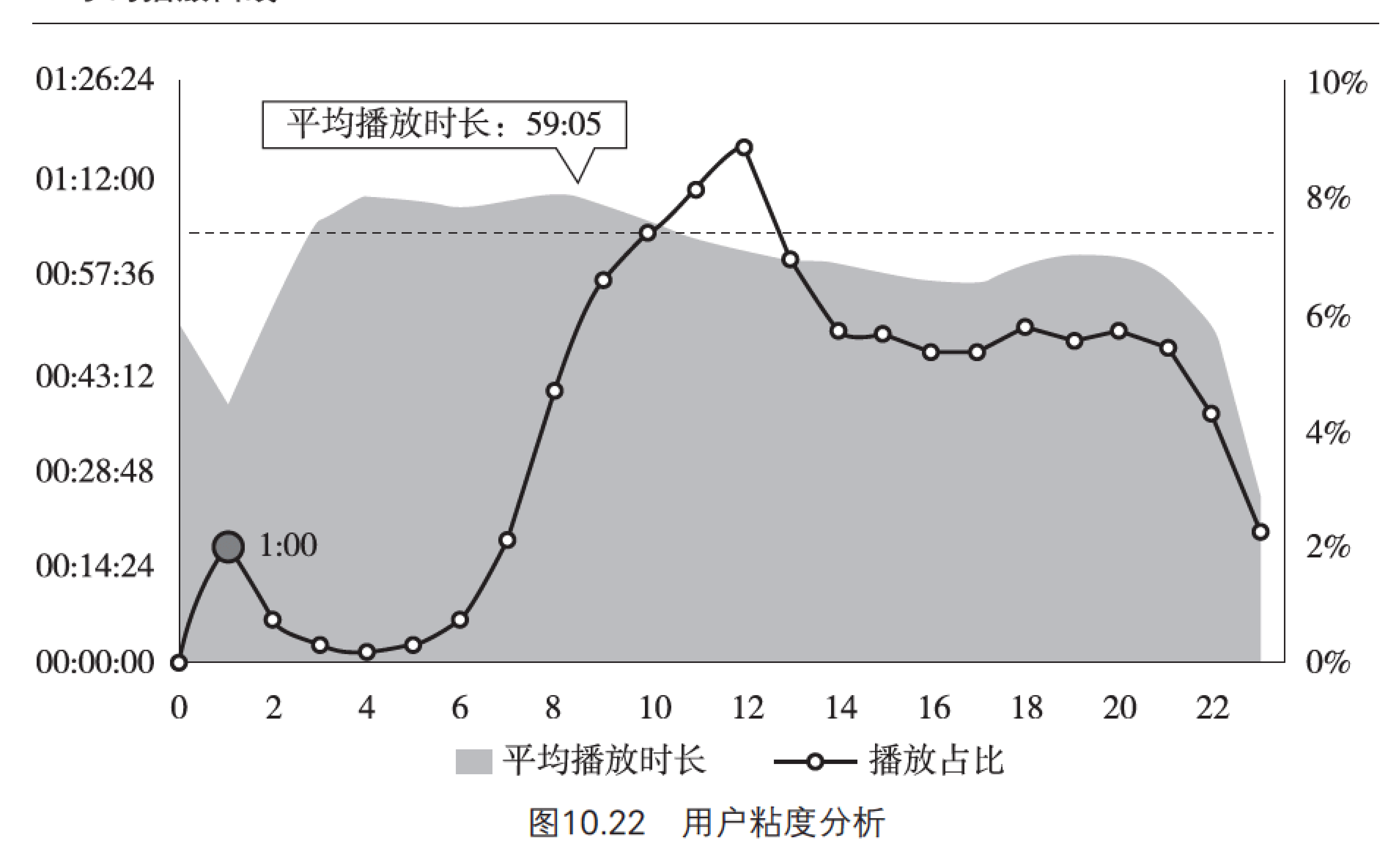


图10.22　用户粘度分析

从用户回看状态曲线可以看出观众对哪些情节特别关注，尤其能反映观众对该情节的喜爱程度。从芒果TV视频用户的回看情况看（如图10.23所示），这期节目有很多亮点，基本上每个时段都有观众回看。

回看状态比例

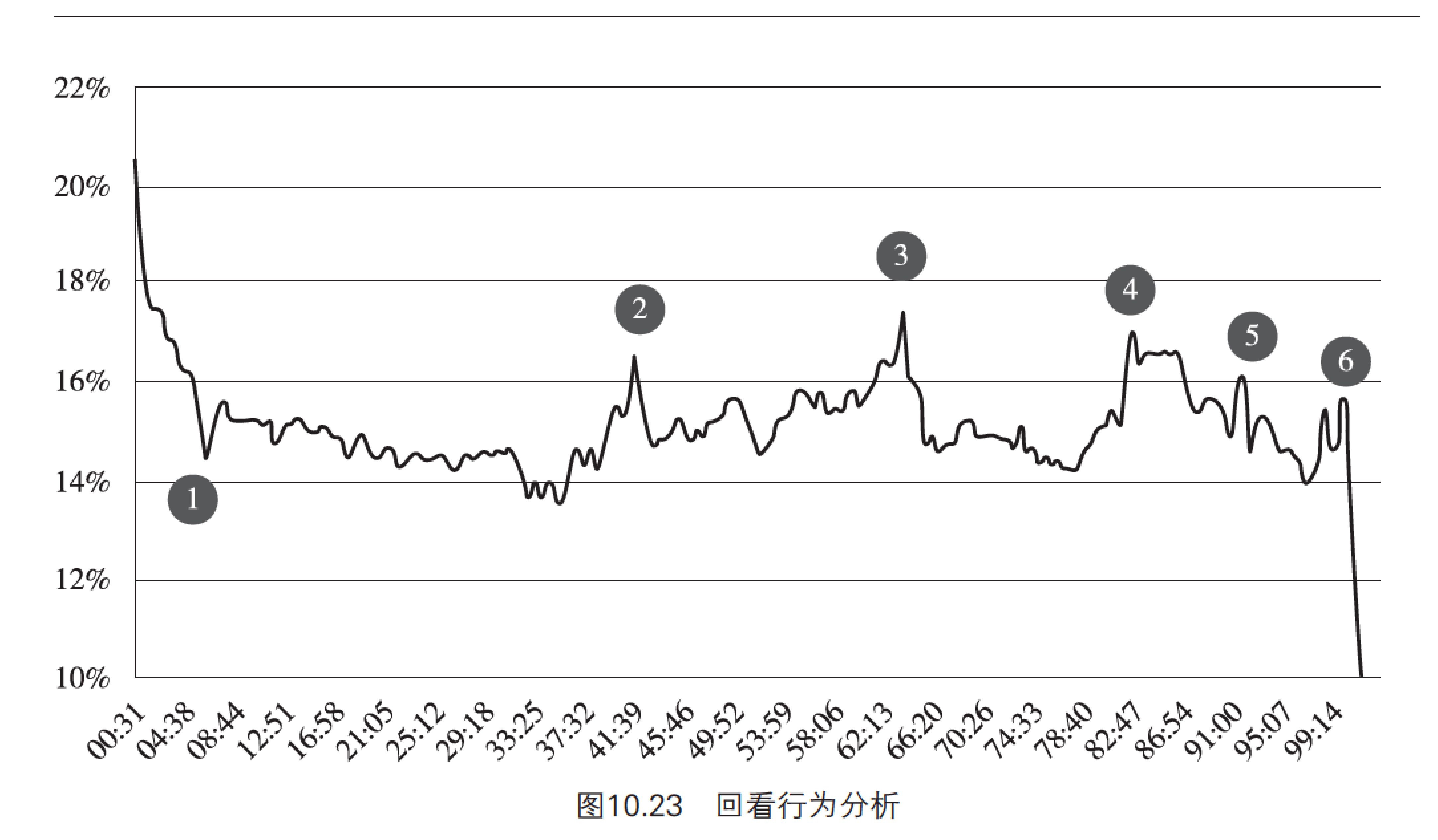


图10.23　回看行为分析

从《爸爸去哪儿》第二季开播以来，微博上网友们对该剧的关注和讨论也空前火热，其中《爸爸去哪儿》话题的阅读量多达33.6亿次，随着节目的播出，阅读量仍在增加。通过大量数据分析显示，有超过60%女性参与讨论，这表明相对于男性，女性群体对亲子节目的喜

爱程度更高。此外，从兴趣偏好角度来看，在男性和女性两类上网人群中，女性网友更关注美食环节，而男性网友则更热衷于科技和体育，如图10.24所示。

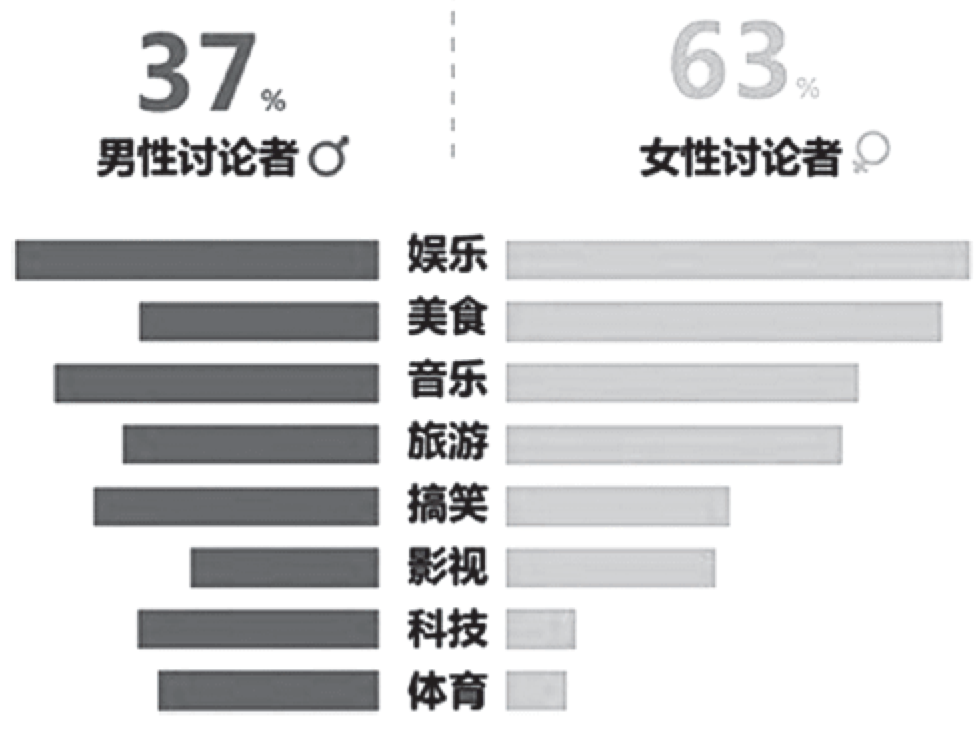


图10.24　微博男女比例及用户兴趣分析

10.4.2　大数据分析电影票房

2014年，汤姆·克鲁斯主演的科幻大片《明日边缘》在北美的票房虽然败给了同档期的爱情片《星运里的错》和由童话故事睡美人改编的《沉睡魔咒》，但此片在海外却有不俗的反响，收获1.4亿美元的海外票房，在亚洲表现尤为突出。著名的娱乐网站Vulture做了一份大数据统计，分析汤姆·克鲁斯在世界各地的票房号召力（如图10.25所示）。此次统计所采取的样本为包括《明日边缘》在内的由汤姆·克鲁斯主演的十部影片。

图10.25　汤姆克鲁斯影片票房总额

从汤姆·克鲁斯主演的影片在世界各地的票房数据分析可以看出，美国人民依旧是其最大的拥戴者，2004年至今，他主演的十部电影一共在美国就收获了超过十亿美元的票房。从榜单上也不难看出汤姆·克鲁斯在亚洲也是极具票房号召力的。中国、日本、韩国三大票房收入毫无疑问全部跻身前五名。

从海外票房占票房总额的比例能更直观地看出，海外市场对汤姆·克鲁斯影片票房起到至关重要的作用，如图10.26所示。在这十部影片中，就有高达九部的海外票房占总票房比重超过50%，最高的《明日边缘》更是接近80%。这样"恐怖"的数据足以显示出汤姆·克鲁斯巨大的国际影响力。

由于各地区的消费水平不同，人口数量也存在差异，如果算人均票房的话，结果又是另一番景象了，如图10.27所示。很明显，中国由于人口基数太大，人均消费票房一下子就被远远甩出了榜单，而冰岛却跃居到第一位。美国排名第三，亚洲地区则是韩国、新加坡位置最靠前，分列第五、第六位。从这份数据也能看出，汤姆·克鲁斯在欧美亚非拉等各个地区都有一定吸引力，观众的地域分布十分广泛。

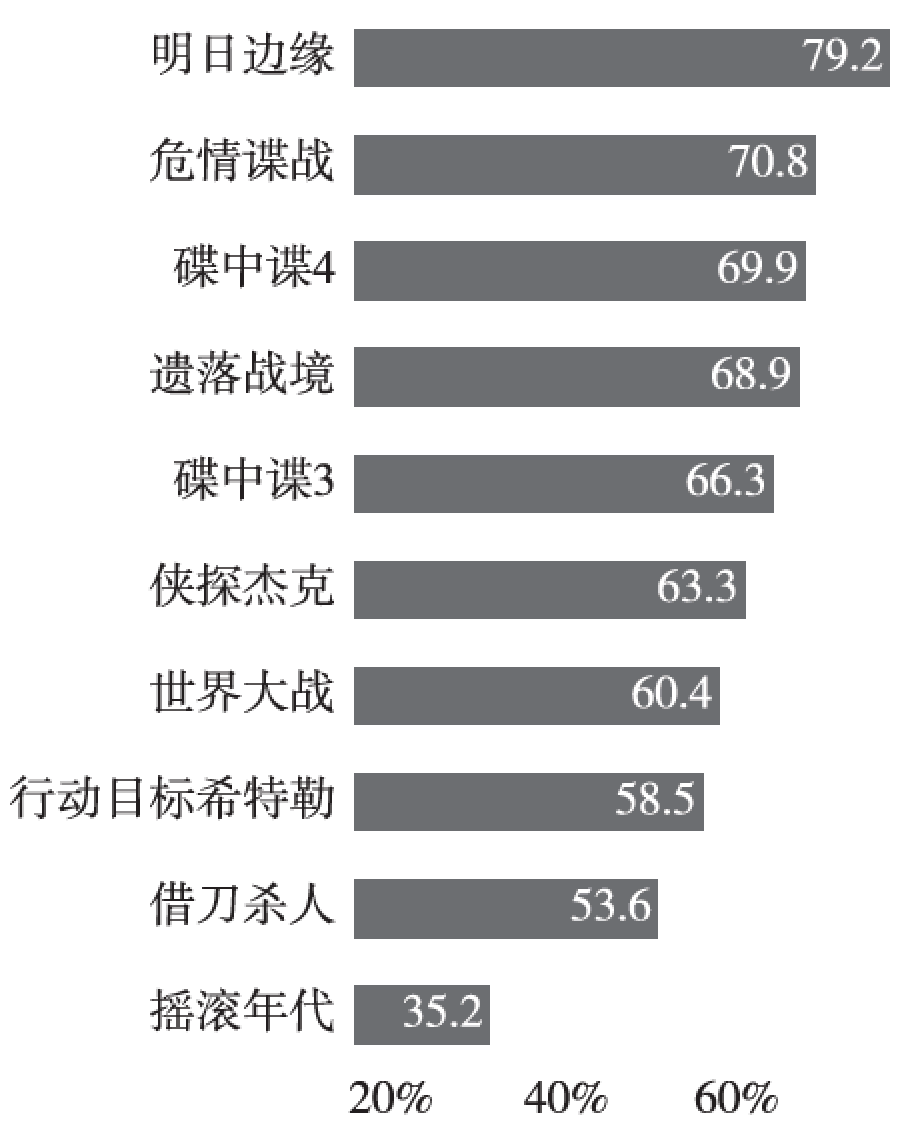


图10.26　汤姆克鲁斯影片海外票房占总票房总额百分比

图10.27　世界各地区人均消费汤姆克鲁斯影片票房

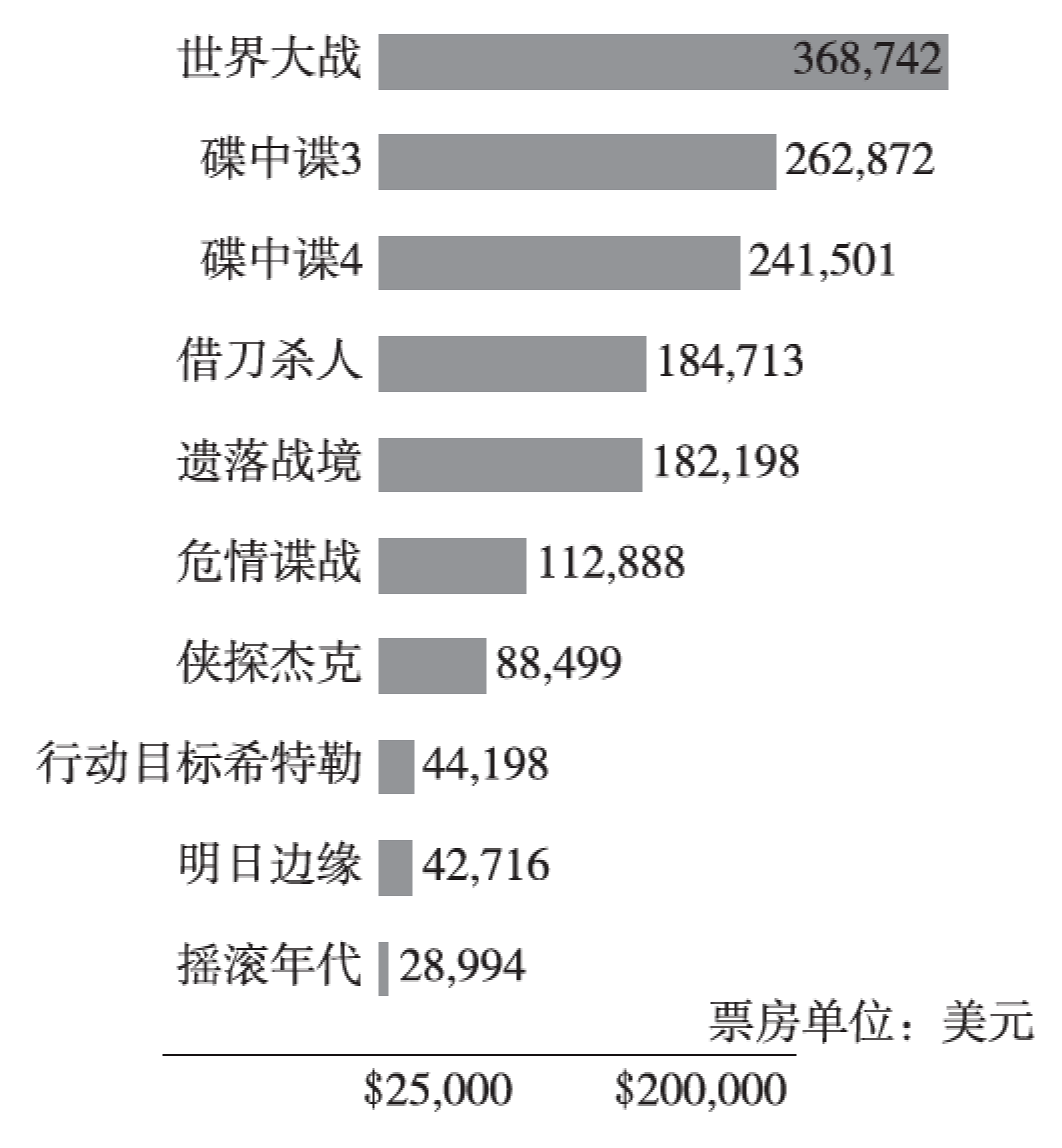


图10.28　冰岛人民最爱的汤姆·克鲁斯影片

接下来把位于第一位的冰岛数据单独调出来，研究看看哪一部影片最受冰岛人的喜爱，如图10.28所示。从数据中一目了然地看出，排名第一的是《世界大战》，比第二名的《碟中课3》高出近10万美元的票房。由此可以看出，史蒂芬·斯皮尔伯格与汤姆·克鲁斯这样的名导＋明星的组合还是很受观众欢迎的。《碟中课3》和《碟中课4》不出意外也分别占据第二、三名。

10.5　练习

1. 归纳大数据在金融行业有哪些具体应用场景。
2. 通过了解大数据在金融行业的应用案例，分析大数据能为金融行业带来哪些价值。
3. 除了本书的案例，举例说明大数据在医疗领域还有哪些应用场景。
4. 分析大数据能为互联网企业带来哪些变革。
5. 思考大数据会对我们的生活产生怎样的影响。

参考文献

[1] 张引，陈敏，廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展. 2013.
[2] 覃雄派，王会举，杜小勇等. 大数据分析——关系数据库和MapReduce技术的竞争、交融与共生[J]. 软件学报. 2011.

第11章 大数据应用的主流解决方案

Hadoop结合了成本低、可扩展性佳以及无需构建预定义模式就能灵活地处理任何数据等优点[①]，是目前大数据解决方案的主流平台，也是顺应未来大数据和云计算环境的平台。与此同时，在成熟的管理和稳定性上，主流市场对其提出了更高的需求，例如实现SQL环境具有的丰富功能等。因此Hadoop厂商在不断地努力开发更加成熟可靠的工具，对其功能进行完善，并不断创新技术。作为开山鼻祖的Cloudera拥有CDH发行版和配套的管理软件，该公司在Hadoop生态系统中规模最大、知名度最高，也是Hadoop软件最主要的来源，它还是为Hadoop提供企业支持和培训服务的最大供应商，力求Hadoop的企业安全性。亚马逊很早就进入了这个领域，在2009年推出了Amazon Elastic MapReduce，对Hadoop的需求和应用可谓了若指掌。Amazon Elastic MapReduce运行在亚马逊简单存储服务（Amazon S3）和亚马逊弹性计算云（Amazon EC2）上，对于数据密集型任务，用户需要多大容量，就能配置到多大容量，是一项能够迅速扩展的Web服务。2011年，MapR和Hortonworks（后者从雅虎拆分出来）向外界宣布了各自的Hadoop软件发行版，并且为客户提供相关的支持和培训服务。MapR的发行版比较独特，它摒弃了不喜欢的，尤其是它认为是单一故障点的Hadoop分布式文件系统即HDFS，把文件系统换成了基于UNIX的网络文件系统，从开源的Apache项目中获取所需要的组件。接着人们认识到大多数Hadoop用户最终希望实现的是数据分析，因此专门针对Hadoop的数据访问、商业智能和分析工具厂商，如Datameer、Hadapt和Karmasphere[②]就应运而生了。Datameer主要是将商业智能用到大数据上。Hadapt主要提供了能够分析操作Hadoop里面的数据，并且能够分析SQL环境中结构化数据的一体化分析环境。而Karmasphere可以运用SQL及其他语言即席查询和进一步地分析，还可以直接访问Hadoop里面结构化和非结构化的数据。2011年，随着五家主要的数据库和数据管理厂商积极地接受了Hadoop，Hadoop开始迈向主流，而IBM、EMC、Informatica、微软和甲骨文也不断加入到Hadoop领域，使得Hadoop领域的竞争愈发激烈。其中IBM和EMC发布了各自的Hadoop发行版，而且EMC还与号称下一代Hadoop的MapR结为合作伙伴。为了支持Hadoop，Informatica扩展了其数据集成平台，将其数据转换代码和解析代码直接融入到Hadoop环境中。而微软和甲骨文则分别与Hortonworks和Cloudera合作，EMC和甲骨文还发布了随时可以运行Hadoop的专门定制的硬件设备。

① http://wenku.baidu.com/view/ebbf351a59eef8c75fbfb3e7.html

② http://blog.sina.com.cn/s/blog_669fa76a01016y5e.html

11.1 Cloudera大数据解决方案

"Hadoop改变了有关数据的整个谈话。"Cloudera首席执行官Tom Reilly说。Cloudera是Hadoop生态系统中规模最大、知名度最高的公司，有100多个客户。该公司致力于大中型企业整体信息化服务、中小型企业门户网站、商业智能、系统集成和应用软件的开发，为企业提供更高效、更安全、更专业的优质IT服务，引导企业进入云计算时代。Cloudera的企业解决方案包括Hadoop软件发行版、Cloudera管理器及支持，其产品和服务如图11.1所示。

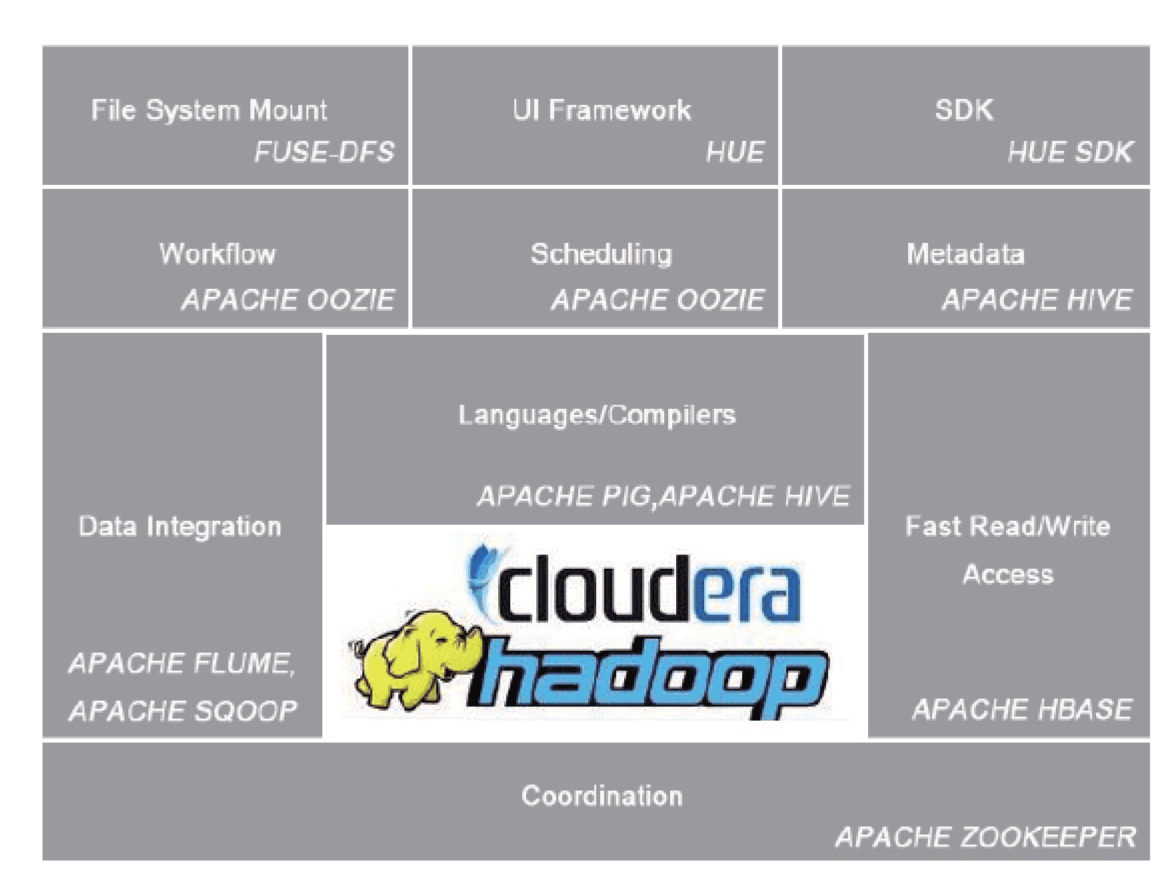


图11.1 Cloudera的产品和服务

为了从数据中获取商业价值而不用担心如何管理Hadoop软件框架，2011年Dell和Cloudera联合推出了Hadoop解决方案Cloudera Enterprise，从而使企业可以更轻松地使用开源的Hadoop。为了帮助大型和中小型企业从数据中获取更多的商业价值，部署基于开源的Apache Hadoop解决方案，Cloudera在其企业分析数据管理软件中，采用了Apache Hadoop驱动，与英特尔至强技术的数据中心架构一起，提供了客户驱动的大数据解决方案。英特尔与Cloudera在大数据创新方面的合作更加紧密，而Cloudera也将致力于开发基于英特尔x86架构的Hadoop应用解决方案。NetApp和Cloudera公司联合发布了Cloudera's Distribution including Apache Hadoop（CDH）和Cloudera Enterprise。Cloudera Enterprise是一项订购服务，由Cloudera Support和Hadoop管理软件组成，可通过NetApp Open Solution for Hadoop加速Apache Hadoop的企业部署和生产应用。CDH为Apache Hadoop企业应用和生产应用创造了条件，它有助于深入了解Hadoop集群，还能够自动执行保持和提高运行质量所需的持续系统变更。

Cloudera公司和NetApp的合作为企业提供了一款专有的开放式解决方案，该解决方案具有高度的可扩展性和企业级存储功能，可显著提高性能并降低成本。合作客户能够利用Hadoop提高分析应用的使用率，从密集型数据和高计算负载中获得实时的结果。通过合作，NetApp和Cloudera提供了开放式的解决方案，从而实现了一流的合作伙伴解决方案组合。

近日，Cloudera收购了大数据加密专业厂商Gazzang，其竞争对手Hortonworks公司在不到一个月之内收购了新兴的安全企业XA Secure。这可能预示着安全已经成为不容忽视的核心问题，而此次Cloudera收购Gazzang将使其Hadoop解决方案在安全性方面获得企业级市场的青睐。

11.2 Hortonworks大数据解决方案

Hortonworks是由Yahoo和Benchmark Capital于2011年7月联合创建的一家企业管理软件公司，总部设在加利福尼亚州。该公司专注于Apache Hadoop框架，支持跨计算机集群分布式处理大型数据集。该公司完全支持开源的软件，其所有的代码都会回馈给Apache Hadoop项目。

主要产品是一款开源的基于Apache Hadoop的数据分析系统Hortonworks数据平台（HDP），如图11.2所示。该平台是专门用来应对多格式和多来源的数据，处理起来更简单、更有成本效益，除此之外还提供大数据云存储，大数据处理和分析等服务。HDP为了更容易地集成Apache Hadoop的数据流业务与现有的数据架构，还提供了一个稳定、开放和高度可扩展的平台。该平台包括各种的Hadoop分布式文件系统（HDFS）以及Apache Hadoop项目、MapReduce、Pig、Hive、HBase、Zookeeper和其他各种组件，使得Hadoop的平台更易于管理，更具有开放性和可扩展性。图11.3展示了HDP的数据来源、数据类型以及应用。

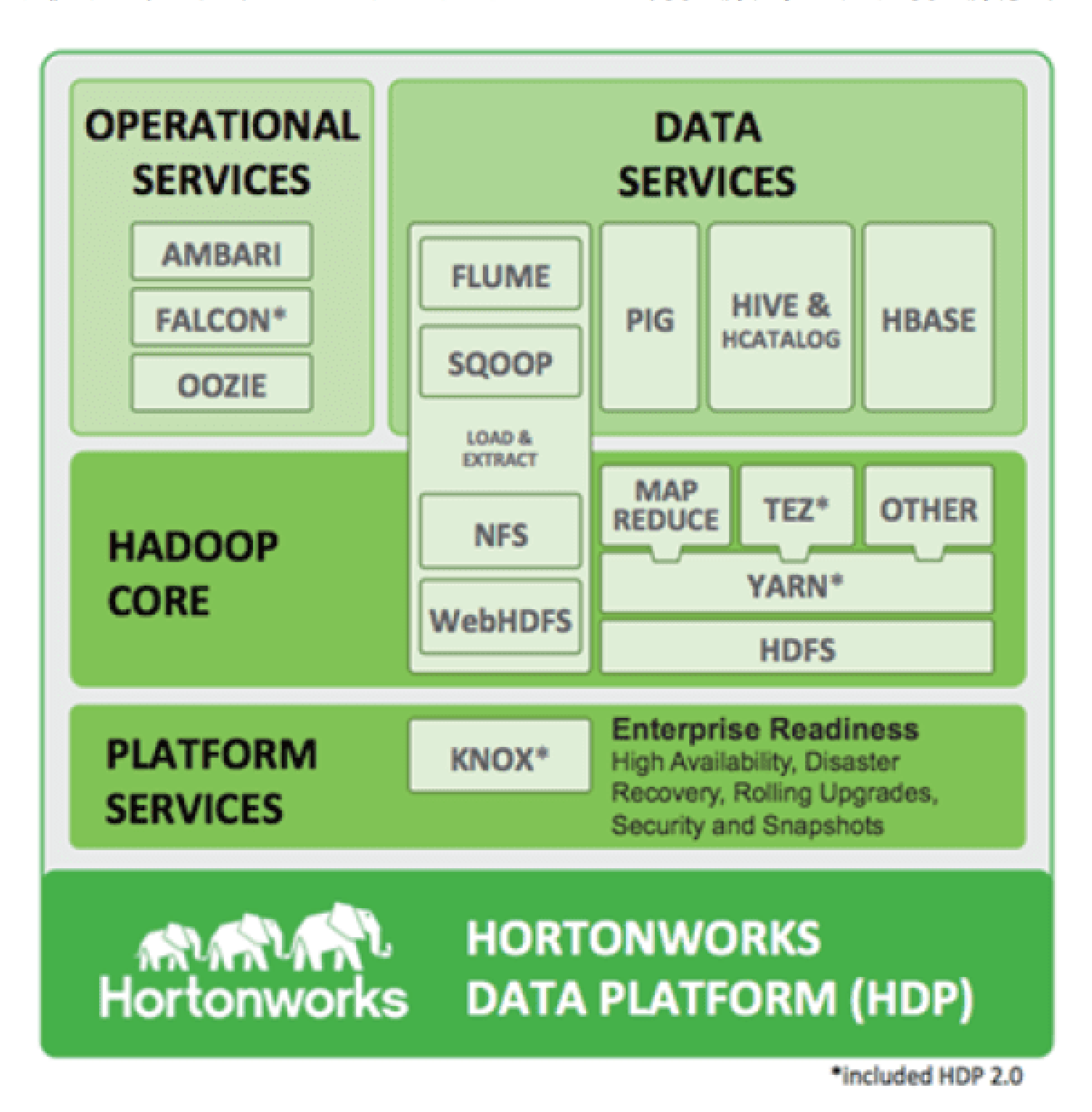


图11.2　HDP（Hortonworks数据平台）

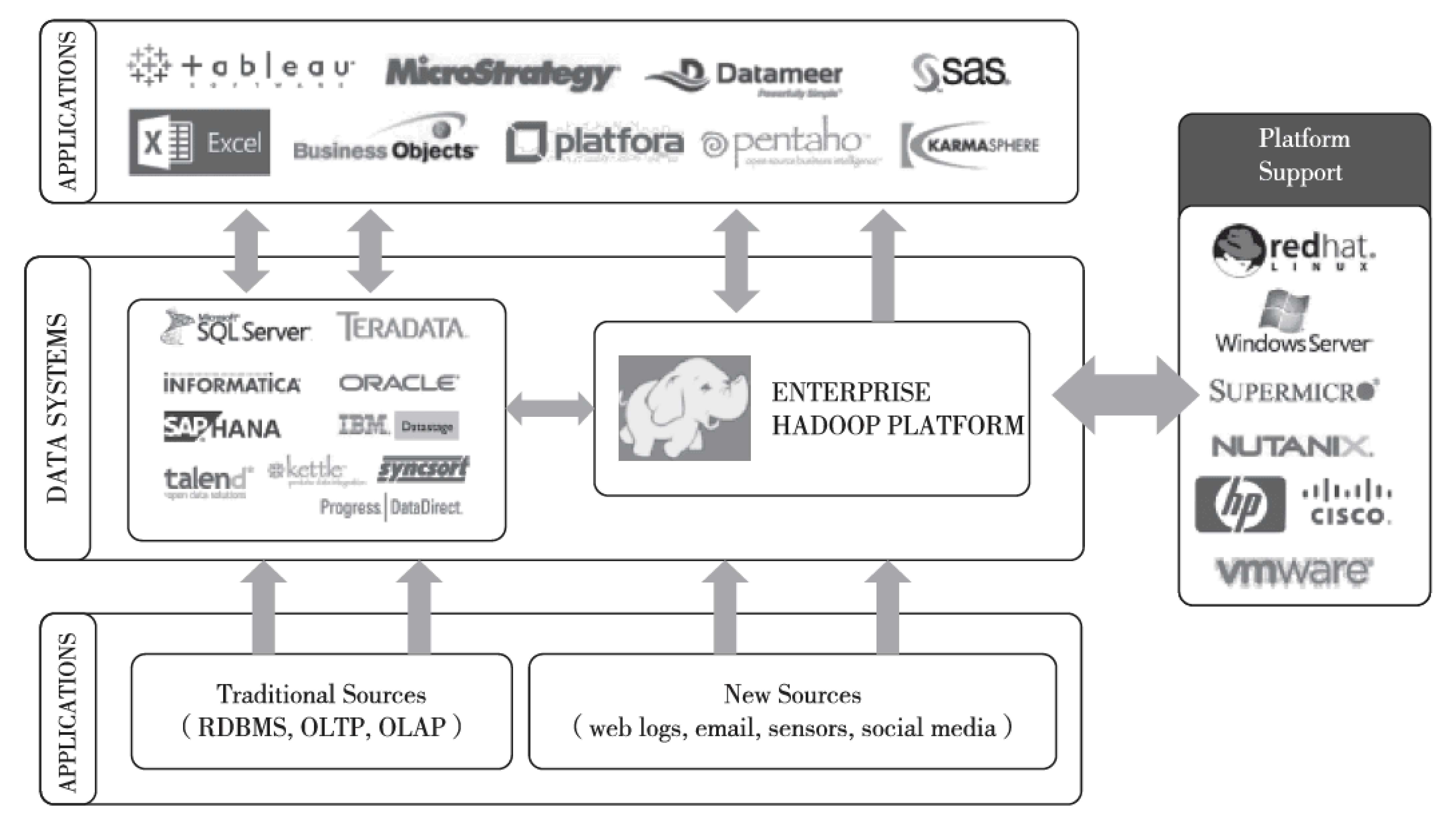


图11.3　Hortonworks HDP生态系统图

Hortonworks还提供Hadoop支持、咨询和培训，在竞争力度上与Cloudera和MapR不断加大。该公司的合作伙伴已超过140个，其中包括微软、Rackspace和Teradata等行业巨头。在与微软的合作中，Hortonworks将帮助微软开发一款遵循Apache开源项目原则的、与Windows兼容的Hadoop版本；与Rackspace的合作旨在Hortonworks自己的云服务中实施Hadoop服务；而与TeraData的合作则旨在帮助企业建立基于Hadoop的大数据分析环境。

11.3 MapR大数据解决方案

MapR公司是美国加州的圣何塞市的一个企业管理软件公司，它使Hadoop变成一个速度更快、可靠性更高、管理更容易、使用更加方便的分布式计算服务和存储平台，同时性能也在不断提高。MapR号称是下一代的Hadoop，主要专注于可用性和数据安全优化及开发、销售Apache Hadoop的衍生软件。该公司对Apache Hadoop的主要贡献有：HBase、Pig（编程语言）、Apache Hive以及Apache ZooKeeper。MapR公司的Apache Hadoop发行版提供了完整的数据保护、无单点故障等功能，这大大提高了其性能与易用性。MapR还被亚马逊云服务选择为亚马逊弹性云EC2的升级版本。

MapR将极大地扩大了Hadoop的使用方式和范围。它包含了开源社区的许多流行的工具和功能，例如HBase和Hive，它们完全和Apache Hadoop的API兼容。它还能为客户节约一半的硬件资源消耗，通过海量数据的分析提升竞争优势。MapR目前有两个版本，M3和M5，其中M3是免费的，M5为收费版（有试用期）。Canonical公司副总裁Kyle MacDonald表示[①]，MapR M3是易于部署的企业级Hadoop解决方案，“我们为Ubuntu客户提供了高效执行大数据的新途径”。作为MapR的合作伙伴，EMC采用了M5作为其EMC Greenplum HD企业版的基础。而整个MapR的核心是其分布式NameNode[②]（如图11.4所示）。分布式的NameNode又称Container，与Hadoop原始设计中的NameNode不同，Container不仅维护用户文件的元数据，也维护数据块。每个Container的大小在16 GB～32 GB之间（这也就意味着一个节点上会有很多个Container），同一个Container在不同节点间有副本。

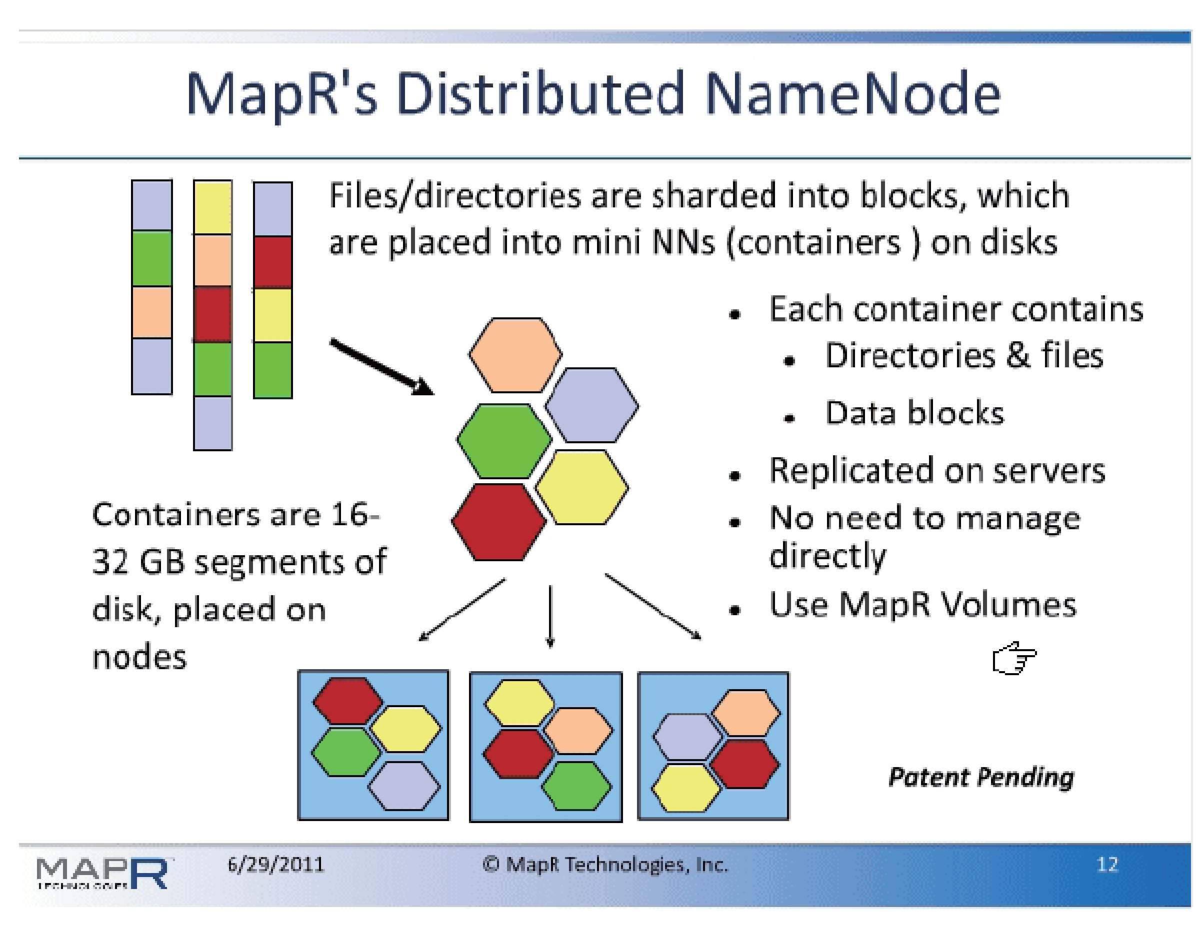


图11.4 MapR's Distributed NameNode

MapR公司的首席执行官John Schroeder表示：“MapR通过为Hadoop用户提供专业咨询服务来获取收入。目前公司大约一半的客户是传统的Web和基于云计算的公司，而另一半则是金融、电信和制造公司。我们希望为我们的客户

① http://mobile.163.com/13/0331/09/8R9MILIK0011671M.html

② http://blog.csdn.net/zhangxinrun/article/details/7335919

提供最好的技术。几乎所有的MapR客户（92%）主要的花费在许可证上，而不是配套服务和支持。”由于MapR认为传统的两阶段提交和基于Quarum的协议（例如Paxos）都有局限性，于是提出了新的解决方案：MapR lockless transaction。

11.4　亚马逊大数据解决方案

亚马逊以企业云平台闻名于世[①]，也推出过一些大数据产品，例如基于Hadoop的Elastic MapReduce、DynamoDB大数据数据库以及能够与Amazon Web Services顺利协作的Redshift规模化并行数据仓储方案。

Amazon Elastic MapReduce（EMR）[②]是一个用于开发专业性较强的应用程序的工具，它使用开源的Hadoop，以便分配数据到一个亚马逊EC2实例集群中。

如图11.5所示，Amazon EMR能够自动加快MapReduce框架在Amazon EC2实例上的Hadoop部署速度，对任务流程中的数据进行细分，使之成为更小的数据块，以便可以并行处理（即“分区映射”功能），然后重新组合处理过的数据，形成最终解决方案（即“规约分区”功能）。其中的数据分析源以及最终结果的输出目的地为Amazon S3。EMR能根据所需配置容量大小，进行数据密集型应用计算，完成Web索引、数据仓库、数据挖掘、日志文件分析、机器学习、财务分析、科学模拟和生物信息研究等任务。Amazon EMR技术专注于数据分析，使得用户无需担心所依靠的计算能力以及费时的Hadoop集群设置、管理或调整。

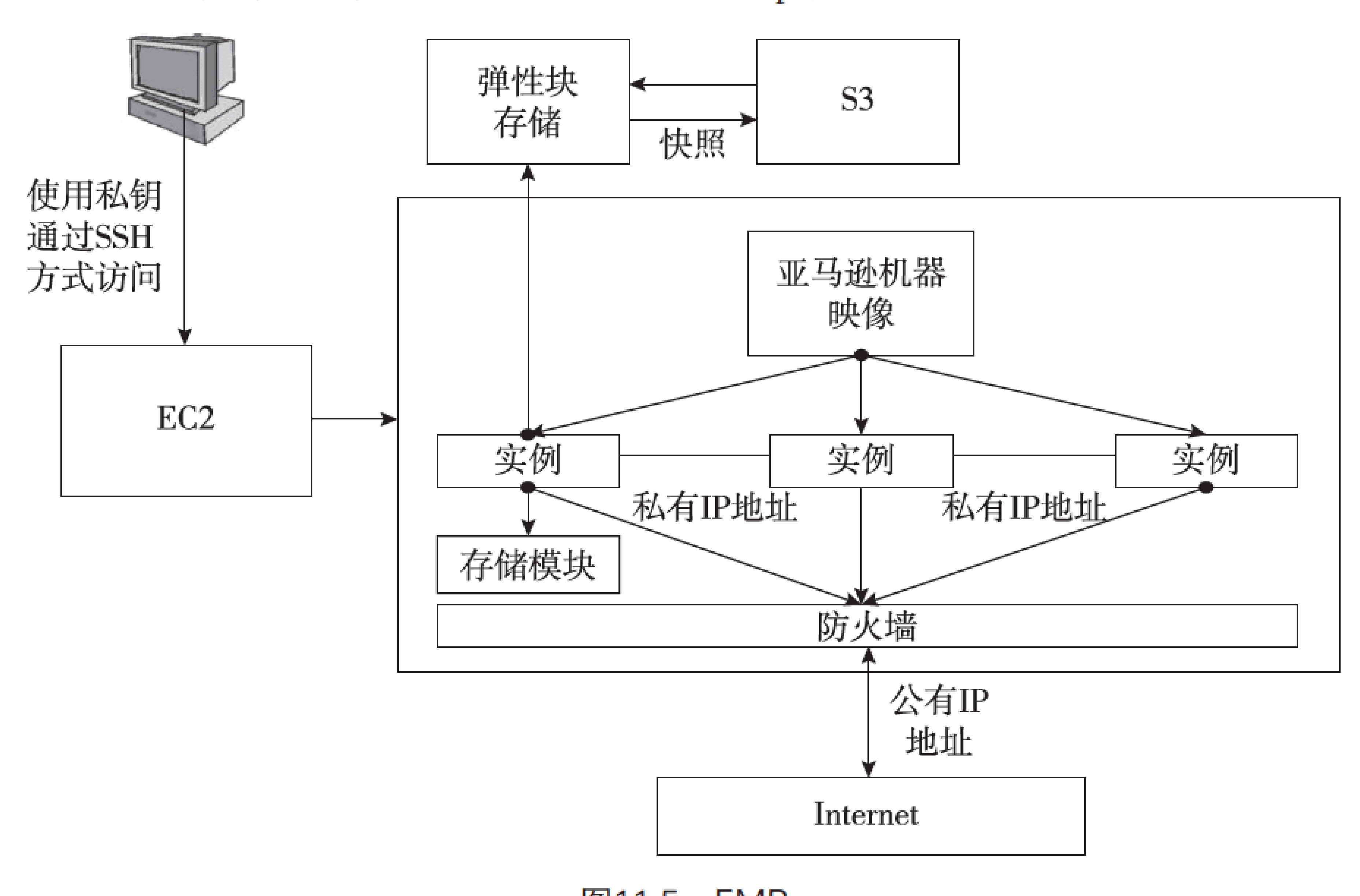


图11.5　EMR

Amazon EMR的特点如下。

- 访问安全。Amazon EMR可自动配置Amazon EC2防火墙，保证实例之间的访问安全

① http://www.techweb.com.cn/data/2013-09-05/1322160.shtml

② http://www.caecp.cn/News/News-1248.html

以及运行任务流程的实例的网络访问安全。用户可以在Amazon Virtual Private Cloud（Amazon VPC）中启动任务流程，通过指定IP范围来隔离计算实例，并使用行业标准加密IPsec VPN策略连接现有的IT基础设施，以确保网络安全。

- 支持第三方工具。为了方便用户开发，Amazon EMR能完美地支持众多的第三方工具和解决方案。
- 应变灵活。Amazon EMR能够运行任意数量的Hadoop计算实例，可用一个、数百个甚至数千个实例来处理数GB、数TB甚至数PB的数据。
- 经济实惠。支付费用低。经过优化的Amazon EMR，可对任务流程的进度进行监控，停用完成流程所占用的资源，尽力为用户节省每一笔开支。
- 使用便捷。Amazon EMR拥有工具和示例数据处理应用程序，无需编写任何代码，提供处理Hadoop集群设置、运行和性能优化。
- 服务全球。Amazon EMR服务使用的EC2基础设施遍布全球。
- 与其他AWS服务集成天衣无缝。Amazon EMR与其他AWS服务（例如Amazon S3、DynamoDB和EC2）的集成为数据处理应用程序提供了坚固的基础设施保障。该服务在Amazon EC2中运行任务流程，在Amazon S3或Amazon DynamoDB中存储输入和输出数据。
- 服务可靠。Amazon EMR是基于Amazon高度可靠的基础设施构建而成的，并且针对Amazon的基础设施环境优化了Hadoop的性能。该服务能监控任务流程执行、重试失败任务、关闭出现问题的实例、配置新节点替换故障节点等情况，以确保任务流程流畅地执行。

2012年亚马逊正式推出了DynamoDB①，它是一款NoSQL数据库产品。DynamoDB为互联网的大数据问题提供了一种快速、可靠且成本低的解决方案，扩展了亚马逊的网络服务（Amazon Web Services）。它结合了NoSQL与云服务，延续了亚马逊上一代NoSQL数据库Dynamo及其基础原理。通过DynamoDB，开发者只需花费较低的成本租用一定量的空间，便可以推广应用，并且随着推广的深入，还可以根据具体规模无限量地扩展容量。Amazon DynamoDB将数据保存在固态硬盘（SSD）上，并且进行跨分区的同步复制，以保证其高可靠性和数据持久性。此外，DynamoDB会在后台将特定数据表的数据和流量分布到各个服务器上，保证客户端平均延迟在10毫秒以内。

DynamoDB特点如下。

- 快速。首先，DynamoDB依托固态硬盘（SSD），使得数据存取速度得以提高；其次，为降低读写操作的延迟，DynamoDB没有为所有属性建立索引；最后，DynamoDB的延迟是由数据存储的分布式特征和请求路由算法决定的，所以可以预测。综上可看出DynamoDB具有低延迟性。
- 灵活。由于DynamoDB的数据模型不是特定或一致性的，因此客户可以根据情况灵活地选择访问方式。此外用户还可以利用DynamoDB提供的原子的递增/递减计数器功能。

① http://tech.it168.com/a2012/0130/1304/000001304459.shtml

- 便利。DynamoDB是完全托管的数据库，开发者可以避免软硬件配置的束缚，借助亚马逊提供的云服务解决数据库从装配到扩展所遇到的一系列问题。
- 低成本。根据DynamoDB的官方定价，读操作为每50个单位容量$0.01/小时，写请求每10个单位容量$0.01/小时，存储数据每$1/月。用户还可以从免费级别的Amazon DynamoDB开始使用，每月可以免费提供40000000个请求。对于创业者来说，云服务要比制备一套软硬件系统所花费的成本低很多。

除上述特点之外，DynamoDB还具有性能高、持久性、可用性以及可预测性等特点。亚马逊首席技术官Werner Vogels表示Amazon Dynamo是在多年经验的基础上创新出来的数据库服务。不过，他也提出还要进一步地验证DynamoDB的具体功能和特性。

Amazon Redshift是一种PB级的数据仓库服务[1]，具有完全托管、简便、快速、安全、兼容和成本低等特点。为了提供快速的查询功能，它采用列存储技术来改善I/O效率并跨过多个节点平行放置查询。为了使用户能够使用各种常见的SQL客户端，它采用标准的PostgreSQL JDBC和ODBC驱动程序。Redshift的数据加载速度和集群大小与Amazon Elastic MapReduce、Amazon S3、Amazon Kinesis、Amazon DynamoDB或任何SSH启用主机的集成呈线性关系。Redshift还能使预配置、配置和监控等与数据仓库相关的大多数常见管理任务自动化，从而把数据自动、连续、递增地备份到Amazon S3上。

11.5　IBM大数据解决方案

2012年5月，IBM软件集团提出了“3A5步”方法论，如图11.6所示。即掌控信息（Align）、获悉洞察（Anticipate）、采取行动（Act）、学习（Learn）和转型（Transform），强调把数据转换为价值，并推出了整体的大数据解决方案——智慧的分析洞察。由此，IBM在大数据市场的精心布局可见一斑。

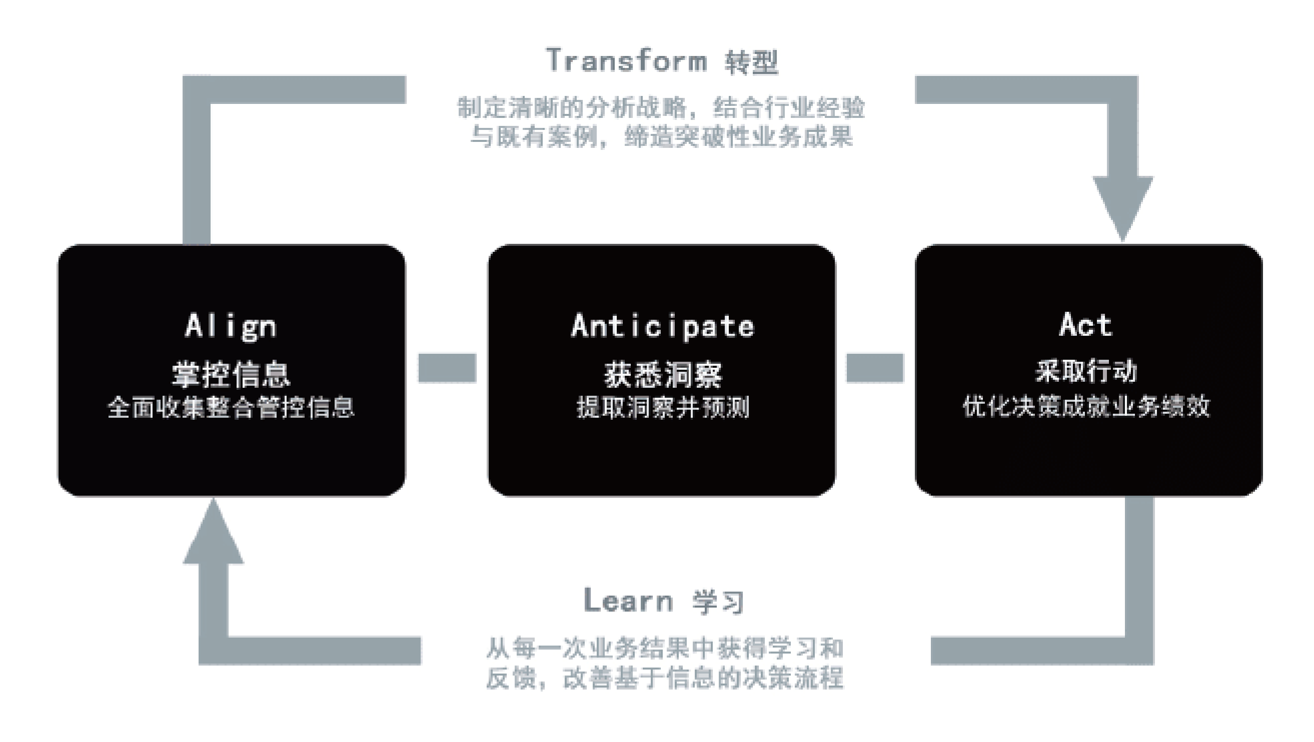


图11.6　IBM“3A5步”模型

图11.7中，IBM的大数据分析平台战略支持与客户现有的系统集成，可以帮助企业解决大数据的挑战，具体包括信息整合、数据治理、元数据管理、Hadoop企业版系统（InfoSphere Big Insights）、流计算（InfoSphere Streams）、数据仓库（PureData System、Infosphere

[1] http://aws.amazon.com/cn/redshift/

Warehouse）、加速器（Accelerator）、可视化和发现（Infosphere Data Explorer）、应用程序开发和系统管理等功能。通过紧密地集成Hadoop与IBM信息管理系统，能够实时对流数据以及非结构化数据分析。

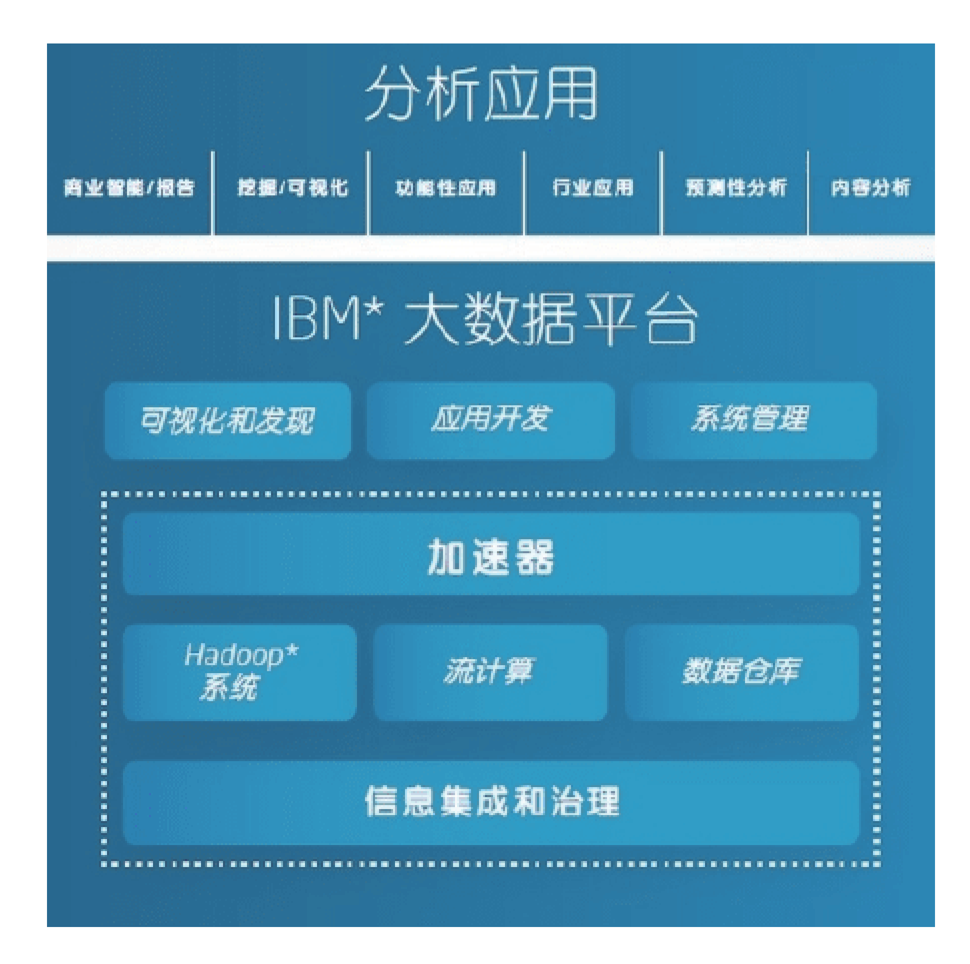


图11.7　IBM的大数据平台

InfoSphere Big Insights是IBM的核心，它包括IBM BigSheets、Apache Hadoop发行版、面向MapReduce编程的Pig编程语言以及针对IBM的DB2数据库连接件。它将Apache Hadoop与IBM的多项创新相结合，提供包括复杂的文本分析、用于数据分析的IBM BigSheets以及一些安全和管理功能，最终得到一款可用于复杂大数据分析的经济高效的用户友好型解决方案。InfoSphere Streams能从几分钟到几小时的窗口中的移动信息（数据流）中揭示有意义的模式。该模式能合并多个流，从多个流中获取新洞察，还能获取低延迟洞察，为注重时效的应用程序获取更好的成果。InfoSphere Streams是IBM针对流计算提供的产品。为提高企业大数据的分析速度，IBM采取了两种途径来解决此问题[①]：一是借助于BLU技术将大数据变成“中数据”甚至是“小数据”；二是对硬件进行优化，推出了面向Hadoop的大数据机。因此，针对第一种途径IBM为提高大数据分析速度发布了BLU Acceleration分析加速技术，用户查询时，BLU快速缩小数据分析范围，清洗海量数据，使得只有小部分有效数据进入分析流程。在提高大数据分析速度的第二种路径——硬件优化方面，IBM发布了IBM PureData box，它是专为Hadoop大数据处理平台设计产生的。PureData Systems大数据专家集成系统被IBM定位为大数据时代的分析处理引擎，使用PureData的用户能在90分钟内完成数据加载。

IBM业务分析软件提供了一套完整的解决方案，主要是为了帮助企业认识业务发展趋势、预测事件和提高绩效，使企业用户可以根据不同的分析结果制定决策，降低风险和成本，提高业绩。它包括五个子产品线，其中Cognos BI平台是一个多层次结构，具体包括以下几个方面。

- 展现层。包括Web、Windows客户端和移动客户端三种。通过Web方式，用户可以访问所有的Cognos BI功能，例如专业报表、记分卡、多维分析、即席查询和仪表盘等，且不需要安装任何插件。Cognos支持与MS Office的无缝融合，支持在移动终端

① http://blog.csdn.net/gnicky/article/details/8773432

上运行，支持iPhone、iPad、Windows Mobile、Symbian、Blackberry等移动平台。

- Web层。主要用于部署Cognos的网关程序，该网关程序可以部署在Apache、IBM HTTP SERVER、IIS或其他中间件上，用户通过浏览器访问时，访问请求会首先发送到网关，网关再发送给BI Server进行处理。
- 应用层。报表统计、即席查询、多维分析、内容管理和内容服务等都被定义为服务，不同服务间通过Cognos BI Bus即不同Service间通信的公共协议进行交互。
- 数据层。包含BI平台支持的各种数据源，例如关系型数据库、多维数据仓库和企业级应用等，Cognos在统一的元数据基础之上支持多数据源。

Cognos BI的基础是Frameworks Manager，负责对来自数据集市（或关系型数据库以及应用系统的数据源）的数据结构进行建模，在这些数据模型的基础上进行即席查询和报表统计的开发，也可以基于这些模型进一步进行OLAP多维分析建模（使用Transformer）并最终生成PowerCubes数据立方体，基于PowerCubes可以进行各种应用程序开发，用户可以通过Web浏览器访问最终的应用程序。

近日，IBM在华发布了最新的Power 8技术，并基于Power 8推出了一系列Power Systems服务器。全新一代的Power服务器是IBM在OpenPower联盟之后推出的第一款处理器产品，围绕大数据分析负载进行了优化。与自营模式相比，OpenPower采用协作和开放的创新模式，将全球顶尖的Power技术开放给业界共享，这将加速系统科技的创新速度，为全球IT产业创造新的增长机会，从芯片、I/O、固件、整机到软件的产业链条的各个厂商都可以利用开放的领先技术获得巨大的商业机会。IBM将通过与OpenPower基金会的合作，打造多方共赢的合作关系和产业环境，更好地满足今天迅速增长的云计算、大数据的需求。IBM全球高级副总裁Tom Rosamilia表示“这是高端服务器技术数十年来第一次真正具有变革性意义的进步，在突飞猛进地改变系统技术、面向新兴应用的同时全面支持一个开放的生态系统，这将帮助客户平滑过渡，成功应对这个数据量和复杂性激增的世界。如今，数据中心的扩展不再适合采用一种放之四海而皆准的方法。通过我们与OpenPower基金会的合作，IBM的Power8处理器将会成为大数据、云计算、移动和社交时代的首选计算平台。”

11.6　甲骨文大数据解决方案

甲骨文（Oracle）的大数据平台是“大数据机、Exadata和Exalytics”三驾马车的组合。甲骨文概括了大数据平台行为的三个方面：数据获取、组织和分析，如图11.8所示。甲骨文为三个阶段开发了不同的产品，并且这些产品与大数据机相融合。大数据机用来捕获数据，再利用Exadata进行分析，然后Exalytics加速BI分析并决策。大数据机是一个软、硬件的集成系统，该系统使用的操作系统是Oracle Linux，配备有Oracle NoSQL数据库社区版本和Oracle HotSpot Java虚拟机。它融合了Cloudera公司的Distribution Including Apache Hadoop（Apache Hadoop发行版）、Cloudera Manager（管理器管理控制台）、甲骨文NoSQL数据库和甲骨文—Sun的分布式计算平台以及一个开源的R语言。甲骨文的解决方案还包括连接件，让数据在传统的甲骨文数据库部署环境或甲骨文内存计算平台Exadata和大数据机之间传递。

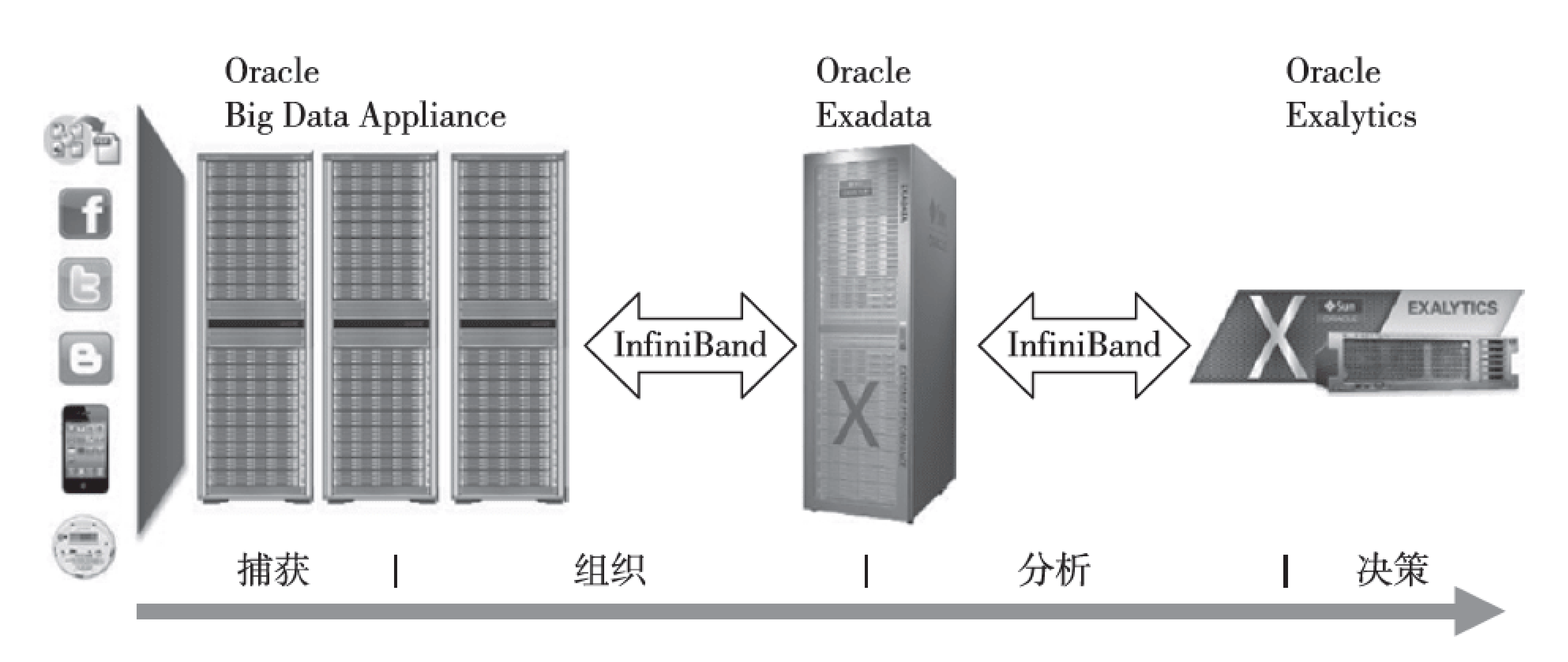


图11.8 Oracle的大数据解决方案

Oracle大数据机具有如下特点：提供了一个使用R分析原始数据源，以及可组织、处理和分析Hadoop中的大数据的平台；为管理海量数据提供了一个具有高可用性的可扩展的系统；完善企业数据仓库并控制IT成本，将所有软、硬件组件预集成到单一的大数据解决方案之中。

甲骨文公司数据仓库技术副总裁CetinOzbutun表示，“Oracle大数据机同Oracle Exadata数据库云服务器、Oracle Exalytics商务智能云服务器和Oracle Exalogic中间件云服务器一起组成了Oracle最广泛的、高度集成化系统产品组合，可以帮助客户获取和管理各种类型的数据，并且与现有企业数据一起进行分析，获得新的见解，从而在充分获取信息的情况下作出最恰当的决策。”当Oracle大数据机、Oracle Exadata数据库云服务器及Oracle Exalytics商务智能云服务器结合在一起使用时，Oracle提供了业界惟一的全面架构，能够减少数据的移动，同时能够存储、管理和分析所有形式的结构化和非结构化数据。借助于最新版本的Oracle大数据机以及最新升级的Oracle NoSQL数据库和Oracle大数据连接器，Oracle可以把这些大数据技术和数据仓库集成在一起，并为客户提供最新的大数据升级。相比于企业自建Hadoop集群，Oracle大数据机可为客户节省高达39%的成本。利用Oracle大数据机增强的软件功能和增加的存储容量，又能够进一步降低客户的总体拥有成本。此外，Oracle作为Apache Sentry项目的共同创立者之一，为Apache Hadoop存储数据提供了细粒度的授权，以展示对企业级大数据安全性的承诺。通过Oracle大数据机，Oracle现在可以提供全面的包括Apache Sentry、预配置Kerberos授权、LDAP授权、强大的集中化审计以及Oracle Audit Vault and Database Firewall在内的Hadoop安全解决方案。甲骨文公司的最新软件产品Oracle Big Data Connectors能使客户利用具有表空间加密功能的Oracle数据库11g，轻松整合存储在Hadoop和Oracle NoSQL数据库中的数据。而配备有Oracle Big Data Connectors软件的Oracle大数据机则借助于Oracle Exalytics商务智能云服务器、Oracle Exalogic中间件云服务器以及Oracle Exadata数据库云服务器，满足客户从企业数据中心获取、组织和分析大数据的所有需求。

11.7 EMC大数据解决方案

创建于1979年的EMC（易安信）是一家美国信息存储资讯科技公司，是全球最大的软件和企业存储设备提供商，主要业务是信息存储、产品管理与服务以及提供解决方案。EMC全

球副总裁兼中国区总裁蔡汉辉指出，“企业用户只需要三步，就可以实现EMC大数据之旅。第一步是搭建云基础架构，EMC给企业用户提供EMC Isilon和EMC Atmos；第二步是进入数据科学协作和自助服务，催生出企业中‘数据科学家’的角色；第三步是实时决策，支持大数据的应用程序，进而实现数据货币化。”

EMC对Greenplum的收购使得EMC Greenplum大数据一体机得以出现。Greenplum这一数据库产品采用Shared-nothing的大规模并行处理（Massive Parallel Process，MPP）架构，对于大数据下的BI等大数据应用有着较好的支持，通过灵活增加节点来实现横向扩展，以便控制成本、提升性能，其大数据战略如图11.9所示。利用大量的并行处理来查询大数据集，Greenplum数据库可以在普通硬件的服务器上运行，对于虚拟化、云计算以及大数据分析，这都是一个非常重要的前提。在BI/DW和全球数据处理领域，Greenplum提供的数据仓库引擎产品和咨询服务是速度最快、容量最大、性价比最好的。目前Greenplum HD Hadoop发行版和传统的Greenplum Database构成了Greenplum的数据库产品，前者可以存储和分析导入Greenplum中的非结构化数据，后者可以用来应对企业的结构化数据。EMC大数据解决方案的核心是由Greenplum HD Hadoop发行版、Greenplum Database以及Greenplum Chorus共同组成的EMC的统一大数据分析平台（UAP）。在这个平台上，数据团队可以无缝地共享信息、协作分析。其中，在行业中处于领先地位的Greenplum Chorus是全世界第一个基于协作分析的大数据平台，是一个社交化的数据处理平台，操作使用习惯与Facebook等网站采用的社交模式很像，它满足了蔡汉辉在第二步中介绍的要求。此外，Greenplum Chorus还可以建立把一定数据变成一个集合的数据沙箱，用户可以利用工具来处理和分析这个集合，对数据的分析结果进行共享。增加数据分析和挖掘的趣味性，使普通用户也可以做到，同时可以交互各种角色，形成一个数据社交圈。

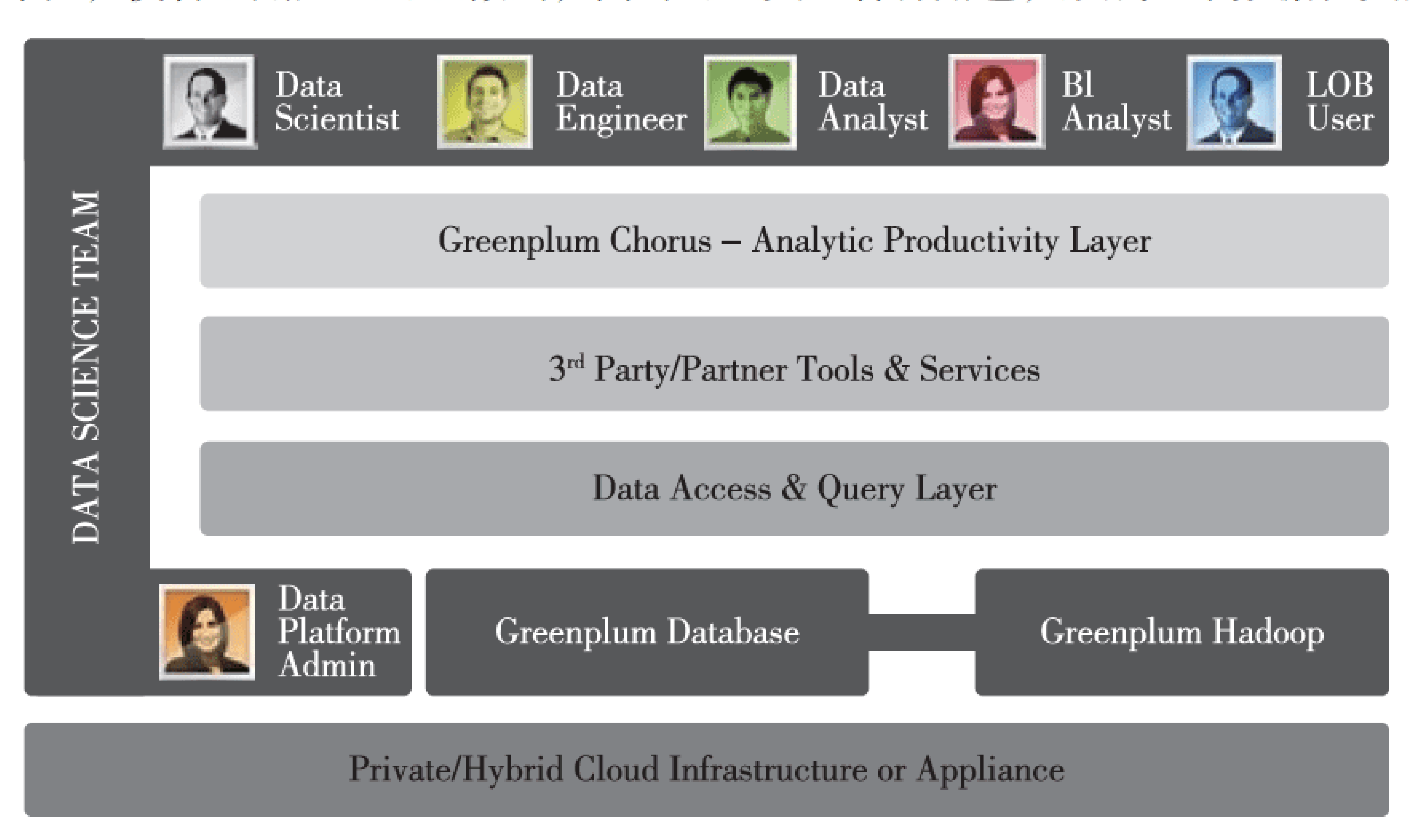


图11.9　EMC Greenplum的大数据战略

EMC Greenplum统一分析平台还包括一个能够将人的智慧和技术产品相结合的“数据科学家计划”。数据科学家能够灵活地利用各种工具去抓取数据，形成数据沙箱，快速地进行实时分析和展现，因此不仅需要具有数据方面的知识背景，还要求具有一定的数学建模能力，同时还要懂得企业内部的运转流程，从而帮助企业将数据变成具有商业价值的信息。

在硬件方面，EMC开发了模块化的，能在一个设备里运行并扩展Greenplum HD节点和Greenplum关系数据库的EMC数据计算设备（DCA）。为了使管理员可以方便地管理、监控和配置Hadoop容量、系统性能以及Greenplum数据库，还为用户设计了一个可以共享的指挥中心（Command Center）界面。另外，随着Hadoop平台日趋成熟与发展，其对分析功能的要求也会急剧增加。

2014年7月9日EMC公司宣布了新一代的Isilon平台和解决方案。它是EMC IsilonOneFS的一项重要升级，是业界首个横向扩展的企业级数据湖，涵盖了对HDFS的不间断支持。通过利用针对数据湖的HDFS，客户得以将Hadoop用于大数据，节省了搬运数据的时间和成本，使得客户显著提升了提取、存储、保护及管理大量非结构化数据的能力。同一天，EMC与Pivotal联合发布了一个新的大数据分析解决方案。该方案以数据湖Hadoop包的形式交付，快速简便，拥有强大的分析能力以及高效横向扩展平台的成本优势，巩固了EMC在Hadoop存储基础架构领域的领导者地位。作为企业级HDFS共享存储市场排名第一的领导者，以及第一个原生集成大数据分析到HDFS的横向扩展NAS提供商，EMC Isilon产品管理及产品市场副总裁Sam Grocott表示，"Isilon横向扩展数据湖是EMC解决全球最大存储挑战的策略，这一挑战就是由传统的和新一代的应用所带来的非结构化数据增长。新推出的EMC Isilon软件、平台以及解决方案旨在帮助客户应对上述挑战，同时驱动更快速的结果产出并节省成本。"

11.8　英特尔大数据解决方案

英特尔（Intel）在大数据时代主要是面向大数据应用[①]，在为存储、网络和云计算提供更高效、更快的架构级别的优化方案方面不断发展；在大数据应用开发，促进软件系统和服务的不断优化和创新方面持续投入；在终端设备和传感器的智能化，构建互联、可管理的和安全的分布式架构方面不断推进。

众所周知，商业价值最突出的大数据处理平台无疑是Hadoop了。英特尔结合自己的硬件技术和成熟经验，凭借在云计算、数据中心领域积累的大量实践经验，以及在开放服务器领域丰富的解决方案，打造了面向大数据应用的Hadoop平台，与其他Hadoop平台相比，该平台能提供可靠性更高、性能更高、功能更多和管理更容易的大数据解决方案[②]。

- 更可靠。全面测试的企业级发行版，能够保证长期运行稳定，用户及时修正漏洞，集成最新开源和自行开发的补丁，保证各个部件之间的一致性，使应用能顺滑运行。
- 性能更高。它能深度结合硬件技术，提高平台性能，实现软硬一体的高效率的大数据解决方案。能够借助Hadoop底层的大量优化算法，使计算存储分布更均衡、应用效率更高。其系统安装程序计算得出的参数配置，适合目前大多数主流平台的应用情况。
- 功能更多。能提供HBase数据库复制和备份功能，提供跨数据中心的HBase数据库虚拟大表功能，此外还有其他针对企业用户的增强功能。
- 管理更容易。能在网页、邮件等发生系统异常时报警。能解决开源版本管理困难的问

① http://mobile.51cto.com/news-392074.htm

② http://server.zol.com.cn/356/3562592.html

题，提供独有的管理界面和集群安装。

Intel的Hadoop企业级发行版构建的基础是开源的Apache架构。通过优化其基础技术层面来提升Intel大数据平台的竞争力。例如，为加速大数据集的分析效率，改进吞吐率和速度，在多核处理能力和高宽带、低延迟等方面进行了优化；为能够处理庞大的数据集，在内存计算、数据压缩和数据保护等方面进行了优化；为给原始数据存储提供高性能、高吞吐率的支持，对SATA接口的固态硬盘（SSD）进行了优化；为加快数据的加密速度和有效保护分布式基础设施，增强了内置于硬件的安全性。

另外，英特尔还推出了发行版Hadoop的免费版。与发行版相比，免费版除了在支持的存储容量和节点数量上不同外，其核心代码和功能都相同。免费版的Hadoop有助于降低大数据应用的门槛，将大数据的Hadoop解决方案惠及更多用户，让更多的用户能够体验Hadoop在大数据处理上的优势。目前，英特尔发行版的Hadoop在电信、生产制造、视频监控等行业都有广泛的应用。Intel为提供大数据应用的指导，帮助推动大数据的应用，提供了大数据应用指南和工具，并提供Intel云构建计划（Cloud Builders）。“英特尔预计，到2015年，将有超过10亿的新用户通过超过150亿的设备接入到互联网中，这一远景无疑将会为技术创新带来独特的机遇，这些机遇涵盖从数据中心到客户终端设备的完整范围。”英特尔数据中心事业部高密度计算业务总经理Jason Waxman谈道，“与行业领袖合作，共同将业经验证的云计算解决方案分享出来，用于满足现今IT所面临的基本需求——这是我们参与云计算的重要方式。截至目前，我们可以看到超过30套通过英特尔云构建计划参考架构交付的业经验证的解决方案。”①

11.9　SAP大数据解决方案

SAP HANA内存计算是SAP的大数据解决方案。它的推出改变了整个市场，在内存计算的发展中所有的软件厂商都在积极地发展着，而SAP HANA将内存计算技术推向了一个更高的位置，它超越了过去的内存计算技术，在该领域成为最先进的领导者。它从一开始就组合并应用了多种技术，在创新软件架构上摆脱了过去的模式，形成多种架构技术的技术长板。SAP HANA是基于开放式架构来设计的，它是一套通过优化软件和整合硬件的基于内存计算技术的应用，是一套灵活、多用途且与数据源无关的基于内存计算的全新平台。承载SAP HANA的硬件供应商可以由企业自由选择，用户可以把它理解成一体机，如图11.10所示。HANA可进行实时数据复制、数据抽取、计算及计划引擎、行/列存储、内存计算引擎以及数据建模。它是一个高性能的实时数据平台，该平台运行在认证的硬件服务器上，并且包含了内存计算引擎和内存数据库。SAP HANA能够通过实时的内存技术来帮助企业提高运营效率②，使客户能够实时分析几乎任何来源的大量数据。通过对现有企业应用系统的计算层进行简化，使企业的业务应用可以直接受益于由此带来的硬件性能的提升。此外，SAP HANA

① http://tech.ifeng.com/internet/detail_2011_04/11/5660070_0.shtml

② http://blog.csdn.net/qinghuawenkang/article/details/9036567

能帮助用户实时浏览和分析任意数据源的业务交易数据，能使企业不断了解变化的大量信息，分析业务的运营情况。在业务发生时，其交易数据将会被实时同步到SAP HANA的内存数据库中，用户可以借助视图来展现分析出来的结果。它还能完成整个企业的扩展性分析，此时只需要在分析模型中添加外部数据即可。

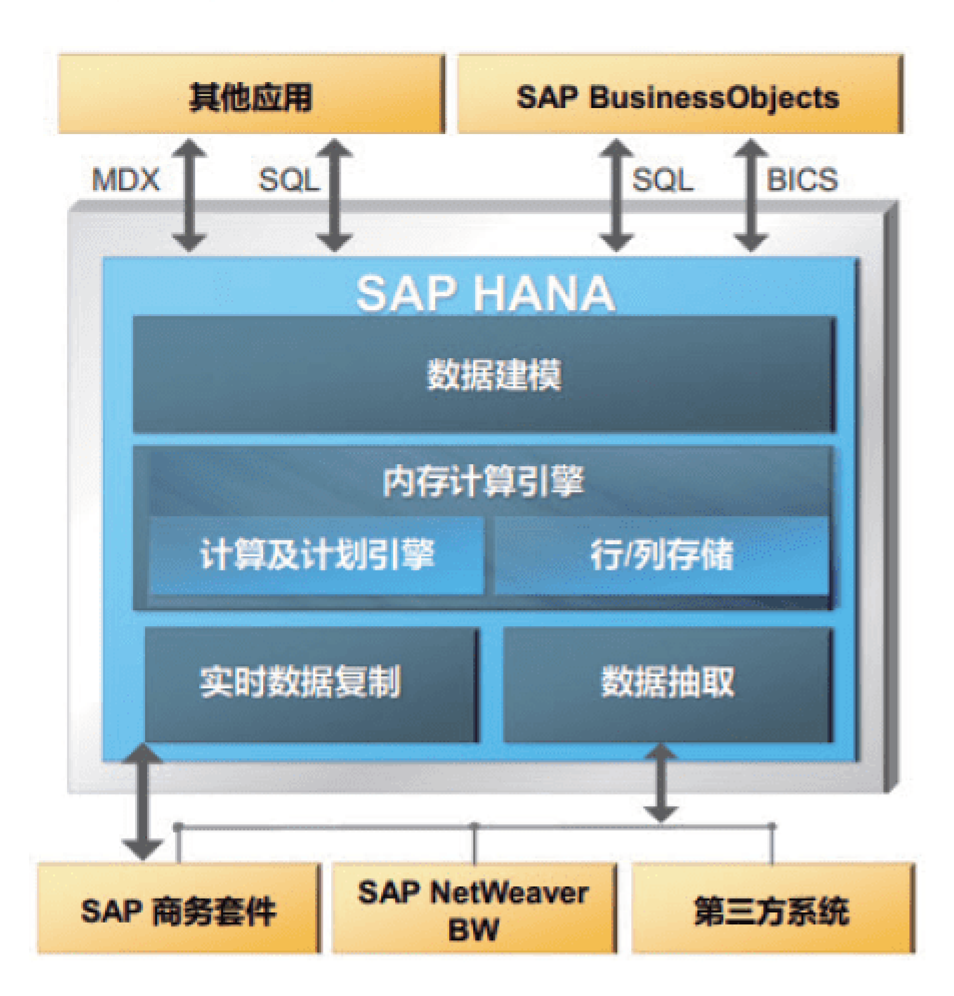


图11.10　SAP HANA技术文档

HANA的内存数据库（SAP In–Memory Database，IMDB）是其重要的组成部分，用来充分挖掘和使用现代多核CPU架构设计所带来的并发处理能力，是包含列存储、行存储和基于对象的数据库技术的一个混合式的内存数据库。HANA的核心是其计算引擎（Computing Engine），支持SQL和MDX语句、SAP和non–SAP数据，负责解析并处理对大量数据的各类CRUDQ操作。包括内存数据库服务器（In–Memory Database Server）、建模工具（Studio）和客户端工具（ODBO、JDBC、ODBC和SQLDBC等）。

HANA的存储结构分为内存、磁盘存储和闪存（持久层）。HANA以内存数据库为基础，还提供了一个2～4TB的闪存来保存内存数据库中的日志信息，能生成保存点和持久层。这是因为磁盘存储的读写速度与内存存在差异，内存中的实时数据的更新或者实时数据同步的操作速度很快。从HANA的内存写到持久层的过程连续不断，中间有一定的时间间隔，其间的数据包含两个部分：数据和日志。其中持久层是HANA内存数据库某个时点的一个完整的镜像备份，以及备份之后在停电前成功执行完毕的所有发生的数据库更新日志信息。因为持久层的容积是有限的，要保存和备份HANA的数据库在磁盘上，所以HANA的备份都保存在外部的物理存储上，比如高速率的硬盘或者其他设备。SAP HANA在内存中进行计算，以下是3种把数据库中的数据复制到内存过程的复制技术。

- 基于触发器的数据同步复制技术（Trigger–Based Replication），根据实时捕捉的SAP ERP的数据库系统的修改变化，几乎实时地同步到HANA中。
- 基于ETL工具的数据复制技术（ETL–Based Replication），通过ETL技术把数据装载到HANA中。
- 基于数据库底层日志的复制技术（Log–Based Replication），根据日志文件把数据库复制到HANA中。

11.10　Teradata大数据解决方案

Teradata（天睿）公司擅长传统数据仓库创建和数据挖掘分析工作，是全球最大的专注于数据仓库、大数据分析和整合营销管理解决方案的供应商。主要的软硬件产品包括：Teradata数据库软件、Teradata逻辑数据模型、Teradata专用平台系列和Teradata分析应用程序和服务。主要的服务包括：Teradata客户支持服务、培训服务和Teradata专业顾问服务。其主要产品和解决方案包括：Teradata数据仓库、Teradata Aster、Teradata统一数据架构（UAD）以及Teradata应用解决方案Aprimo+eCircle。

Teradata公司推出的Teradata Aster大数据探索平台（Teradata Aster Discovery Platform）是业内首个最全面的数据探索解决方案，如图11.11所示。借助于预建的分析功能包，Aster大数据探索平台可以对社交网络、网络点击、客户流失、客户群细分和个性化、情感和传感器数据等进行分析，快速地洞察数据背后的信息。还能通过将MPP数据仓库的优势与MapReduce引擎相结合，为用户提供交互式分析功能，快速挖掘、处理潜藏在数据中的商业价值。Aster大数据探索平台能够做到：快速启用即时数据探索；迭代式开发，完美支持Teradata统一数据架构（UDA）；业务人员使用简单方便。这些使得Aster大数据探索平台与传统的数据探索方式相比具有显著的优势。Aster大数据探索平台包括Aster SQL-H软件、Viewpoint Portlets、Aster管理控制台（AMC）、Teradata Server Management以及海量并行处理5大主要模块，采用Hortonworks Hadoop发行版，提供了基于MapReduce的预封装模块、支持标准的SQL以及数据系统之间高速数据传输的Teradata Aster Adaptor连接器，拥有20多项全新的大数据分析能力，包括可定制的可视化功能等。

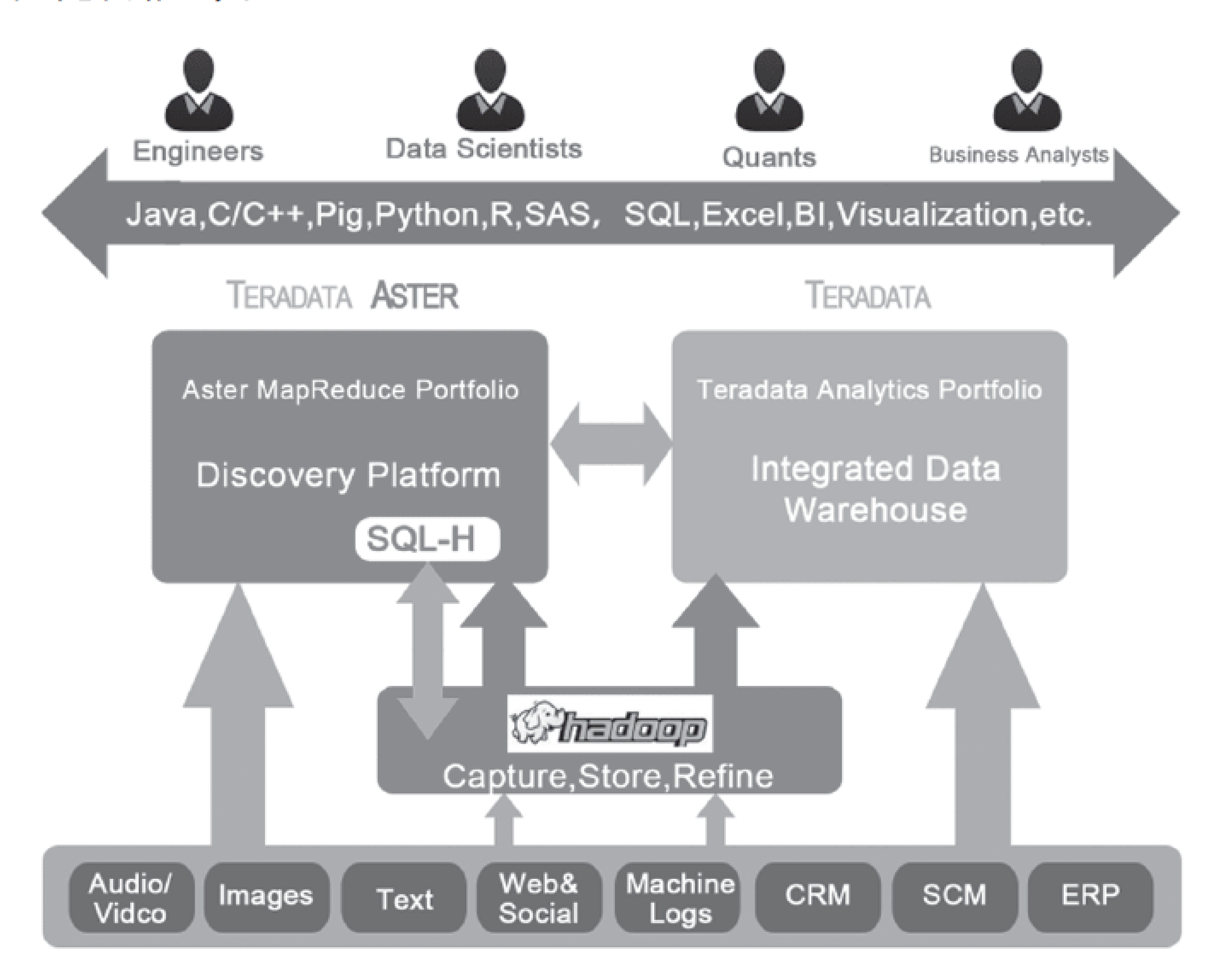


图11.11　Teradata Aster平台

该平台在技术上的重要革新表现在下述几个方面。

- 可视化SQL-MapReduce功能。该平台能够运用数据库中其他的Aster SQL-MapReduce分析功能，以便为客户创建定制可视化功能。

- 数据库内集成Attensity功能。Teradata Aster中集成了Attensity公司在市场上领先的文本和情感分析技术，使得文本和情感分析成为大数据探索过程的重要环节，并能简便地整合到其他分析技术中。
- 借助Zementis，支持数据库内预测模型标记语言（PMML）。该平台集成了Zementis和Teradata Aster，能够让商业分析师通过自选工具来开发统计模型。
- 数据库内集成了“R”执行功能。Teradata Aster数据库内高度集成了常用的开源代码统计语言R。因此，数据科学家可以有更多方式来执行高级分析任务。
- 制造业和金融服务业分析功能。该平台整合了全新的SQL-MapReduce功能，以便更好地支持制造业和金融服务业企业。其中，制造业企业通过使用该功能可以完善工作流程，提高产量并减少浪费，而金融服务业企业通过使用该功能能够更快捷地侦测到欺诈行为和客户流失现象。

2014年5月22日Teradata公司推出了业内最全面的大数据解决方案——Teradata QueryGrid，这也是目前惟一优化企业内外部分析能力的软件。Teradata QueryGrid采用文件系统和分析引擎，使用户专注于数据访问和分析，无需IT人员或专用工具介入。通过在数据的原有存储位置进行处理，最大限度地避免了数据的移动和复制。Teradata QueryGrid还提供了无缝的自助式服务，用户若想访问和分析某个系统的数据，只需在单一Teradata Aster数据库或者Teradata数据库（Teradata Database）查询即可。“Teradata QueryGrid是最灵活的解决方案，配备实现所有功能的创新型软件，得以轻松完成跨数据库分析处理”，Teradata天睿公司实验室（Teradata Labs）总裁Scott Gnau表示，“用户选择相应分析引擎和文件系统后，Teradata软件只要执行一条SQL查询，就能无缝整合不同系统的分析处理能力，无需移动数据。此外，Teradata还支持在单一负载中使用多个文件系统和分析引擎。”

11.11 微软大数据解决方案

在大数据方面微软（Microsoft）公司已经做了很久的研究，并提供了一些解决方案来帮助客户解决大数据带来的挑战。要想挖掘大数据中的价值[①]，首先要收集、存储数据，使数据管理平台可以无缝地存储和处理实时数据及结构化、非结构化等所有类型的数据。微软公司推出的Hadoop发布版HDInsight能够帮助客户通过连接世界的数据和服务来发现新的价值；能够通过与Active Directory、System Center集成，使IT人员可以使用基于企业的安全策略来保护并管理他们的Hadoop集群；能够以Hortonworks Data Platform（HDP）为基础，结合Windows和一个与Apache Hadoop完全兼容的发行版本，可从所有数据中揭示隐藏的知识，并提升企业的洞察力。

微软大数据解决方案将数据和模型与公开的数据服务相结合[②]，使用Windows Azure Marketplace中的应用程序和智能挖掘算法，使用户可以发现更多的数据模式和隐藏的信息。微软的大数据解决方案还能使用企业信息化管理工具，借助SQL Server的分析服务来精

① http://blog.csdn.net/niyi0318/article/details/8157357

② http://www.microsoft.com/china/sql/2012bicampaign/bigdata/

炼数据。

微软的大数据战略包括三个部分：

- 一是对所有类型的数据进行搜集和管理的数据管理层。
- 二是能丰富数据集并将数据变成信息及知识的扩展层。
- 三是能实现数据以及信息的消费化，为从领导层到每个员工提供直观易用的决策支持的洞察力层。

基于SQL Server的微软并行数据仓库一体机使用了Microsoft SQL Server中的“大规模并行处理”（MPP）体系结构及“并行数据仓库”，来提供功能最全面的数据仓库解决方案，是高度可扩展、针对企业数据仓库的设备，是微软大数据战略的重要基础。使用该设备的企业可以完成更复杂的分析、运行更多的报表、处理更大的数据集以及分析更详细的数据，且在运行大规模的查询时速度非常快。此外，它还具有开箱即用的特点，服务器的调整和优化所需时间很少，安装和加载数据的速度也很快，从而减少了工作量，缩短了部署时间，降低了成本。软硬件一体的并行数据仓库一体机只需要很少的调整和优化，可有助于企业降低IT成本。

由于拥有全面的桌面和后端解决方案，微软的大数据战略也被认为是“端到端”的解决方案。微软大数据解决方案有其独特的优势：表现在管理、丰富以及洞察力，如图11.12所示。能随时互连全球的数据，发现隐藏的价值，将内部与公用的数据和服务相结合；前端使用移动终端、笔记本等多种设备，以及微软Office Excel、SharePoint、Power View等工具，从而在数据中获取所需信息，提供决策支持；能借助于支持任何数据的现代数据管理平台，处理任意种类和大小的数据，具有Windows的易用性、云的弹性和可扩展性。

图11.12　微软大数据解决方案的优势

微软全球高级副总裁、大中华区董事长兼首席执行官贺乐斌表示，“微软通过采用先进的算法来帮助用户更高效地挖掘有用数据，再把结果以用户最直观、最熟悉的形式表现出来，从而帮助用户决策。”目前，包括Gartner在内的全球分析师机构，在全球数据仓库方面已经把微软列为主要领导者之一。微软并行数据仓库全球卓越中心总监Russ Cavan表示[①]，诸

① http://www.enet.com.cn/article/2012/1218/A20121218209817_2.shtml

如Powerpivot、Excel、Reporting Services、Analysis Services以及SQL Server Integration Services等表明，微软实现了微软并行数据仓库和多种商业智能工具的紧密集成。微软的大规模数据仓库解决方案不仅能在客户决策时提供灵活、可视化、丰富而且易用的前端展现，同时也能为客户提供后台大规模数据存储、管理与处理，真正实现数据以及信息的消费化。

11.12 国泰安大数据解决方案

深圳国泰安教育技术股份有限公司是一家为金融业与教育业提供综合解决方案的国家级高新技术企业。自2000年以来，国泰安一直致力于为国内外金融业和教育业提供集“研究数据、交易系统、云平台建设、软硬件系统和增值服务”为一体的综合性解决方案，现已发展成为中国教育与金融领域内最专业、最具规模、技术水平领先的综合性解决方案供应商与服务商。

国泰安大数据实验室解决方案从顶层思路的大数据价值链（如图11.13所示）出发，结合市场常用软件以及国泰安自有的软件形成了国泰安大数据实验室解决方案。

图11.13　大数据价值链

如图11.14所示，国泰安大数据实验室解决方案为大数据实验室的建设提供从数据源、大数据采集与ETL、大数据存储、大数据分析与挖掘、大数据展示与可视化五大模块全面系统的服务。首先获取数据，进行大数据采集清洗，然后进行存储处理，分析与挖掘，最后将分析结果可视化展示出来给决策者及相关人员。每个模块下分别配备了相应的软件系统，可以让学生对整个大数据领域的知识与技能进行系统的训练和学习，同时学生还可依托实验室软硬件条件进行相关的研究。

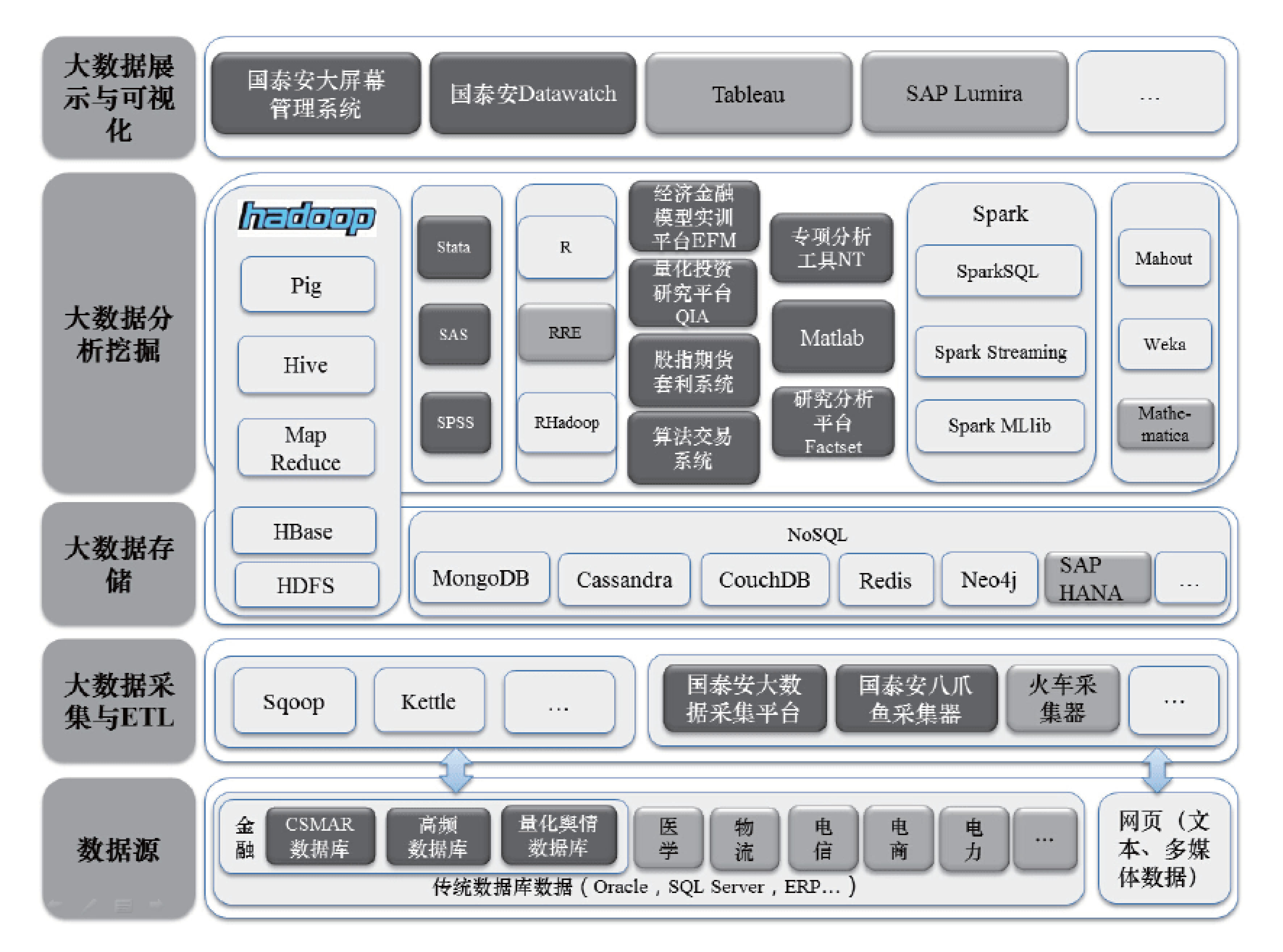


图11.14　国泰安大数据实验室解决方案

其中，国泰安研发或代理的软件用黑色标注，其他公司的收费的软件用灰色标注，其他颜色表示开源或免费软件。

国泰安大数据实验室解决方案（见图11.14）中国泰安研发或代理的产品如表11.1所示。

表11.1　国泰安大数据实验室解决方案部分软件配置列表

软件名称	简　　介
CSMAR数据库	CSMAR数据库是专门针对中国金融、经济领域的研究型精准数据库，包括股票市场、公司研究、基金市场、债券市场、衍生市场、经济研究、行业研究、海外研究和专题研究等11大系列，75个数据库
量化舆情数据库	量化舆情数据库是为了支持新闻传媒、品牌管理和量化投资等研究，通过接收新闻站点、论坛、博客和微博等海量舆情数据而建设的数据存储系统
高频数据库	高频数据库是包含股票、基金、债券、权证、股指期货、商品期货，港交所证券在内的各类高频数据，及基于高频数据传输、更新、应用软件在内的一套整体的系统解决方案
国泰安大数据采集平台	国泰安大数据采集平台实现对各类不同的数据源的手工、半手工、结构化、非结构化和半结构化数据进行统一采集管理
国泰安八爪鱼采集器	国泰安八爪鱼数据采集系统以完全自主研发的分布式云计算平台为核心，可以在很短的时间内，轻松从各种不同的网站或者网页获取大量的规范化数据
Stata	统计分析软件Stata是一套提供其使用者数据分析、数据管理以及绘制专业图表的完整及整合性统计软件
SAS	SAS是一个功能强大的数据库整合平台，可进行数据库集成、序列查询、序列处理等工作
SPSS	SPSS是一系列用于统计学分析运算、数据挖掘、预测分析和决策支持任务的软件产品及相关服务的总称
经济金融模型实训平台EFM	系统提供了200多个主流经济金融财会的模型，并采用步骤式建模的展示方式，将复杂模型分解为简单易懂的模型。基于Matlab的调用强化了编程体验，开源式的模型代码设计为研究者在模型探索中提供了极大的便利。另外，平台接口与CSMAR数据库高度兼容，研究者能够方便调取数据进行实战分析
量化投资研究平台QIA	提供从数据提取及处理、策略构建、策略回验、参数优化到绩效分析整个研究流程的支持，研究完成后的策略可以与iQuant实现无缝对接，进行策略模拟交易和真实交易
股指期货套利系统	该系统具有期现套利、期货跨期套利、组合趋势交易、贯穿整个套利过程的风险监控四大功能
算法交易系统	主要应用于投资机构，使用算法策略进行投资交易。专项分析工具NT：是一款结合GTA丰富的数据库优势以及Quick公司先进的深度分析功能模块，为用户提供精准、及时、富有特色的金融分析终端
研究分析平台Factset	证券分析软件，包含十几种证券分析工具，可进行公司分析、行业分析、固定收益产品分析，全球经济走势预测以及投资组合管理等
Matlab	在数学类科技应用软件中在数值计算方面首屈一指。Matlab可以进行矩阵运算、绘制函数和数据、实现算法、创建用户界面、连接其他编程语言的程序等
国泰安Datawatch	可以实现与多种数据的无缝对接、实时数据连接读取、模型建立和报告生成、显示面板的建立和设置等
国泰安大屏幕管理系统	在显示区域内可任意添加行情展示、技术图表、各类图文、视频影音等多种展示内容，满足了在同一块大屏上实现区域划分，自由划分组合、切换

此外，国泰安针对不同专业量身定做了专门的解决方案，如针对计算机/IT专业更侧重于技术上的学习，可以学习一些分析挖掘上的方法和工具；针对统计/数学专业建议更侧重于拿到行业数据，进行数据分析，建模和挖掘及行业应用的训练；针对金融大数据实验室建议的解决方案则包括了国泰安金融行业数据，以及金融行业相关的一些分析工具等。

11.13 练习

1. 简述目前大数据解决方案的主流平台及其发展要求。

2. Cloudera在Hadoop生态系统中的地位。

3. 基于Apache Hadoop数据分析系统的Hortonworks数据平台（HDP）的主要特点。

4. MapR因为什么提出了新的解决方案MapR lockless transaction？查找MapR lockless transaction相关资料，看其是否解决了所要解决的问题。

5. 找出亚马逊公司的一系列大数据产品的主要特点。

6. 学习IBM的“3A5步”方法论，尝试运用并解决问题。

7. 甲骨文的大数据解决方案包括几个步骤，三驾马车在甲骨文的大数据平台中的作用。

8. EMC大数据解决方案的核心，了解EMC“数据科学家计划”的主要内容。

9. 英特尔的大数据解决方案与其他厂家相比优势何在。

10. 了解SAP HANA内存计算在SAP的大数据解决方案中的重要位置。

11. 查找资料，详细了解Teradata公司在传统数据仓库和数据挖掘分析上的工作。

12. 微软的大数据战略被认为是“端到端”的解决方案，其“端到端”的特点是如何体现的。

13. 查找其他国泰安公司在大数据解决方案上的资料，以便更好地理解GTA大数据架构图。

参考文献

[1] 赵刚. 大数据技术与应用实践指南[M]. 北京：电子工业出版社，2013.

[2] 程永. 智慧的分析洞察[M]. 北京：电子工业出版社，2013.

[3] 刘刚，舒戈. SAP HANA实战[M]. 北京：机械工业出版社，2013.

[4] 胡健，和轶东. SAP内存计算——HANA[M]. 北京：清华大学出版社，2013.

[5] 黄海峰. 蔡汉辉：未来三年EMC大数据业务每年翻番[J]. 通信世界周刊，2012.

[6] 官建文，刘振兴，刘扬. 国内外主要互联网公司大数据布局与应用比较研究[J]. 中国传媒科技，2012.

[7] 徐莲萌. SAP HANA内存计算技术项目实战指南[M]. 北京：清华大学出版社，2012.

[8] 大数据行业发展联合编辑部. 大数据行业内参. 2014：（18）05-06.

[9] MapR新获3000万融资助力Hadoop推广[J]. 计算机与网络. 2013.

[10] 方俊君. 应对大数据挑战. 软件和信息服务. 2013：（60-61）.